A containerised approach to labelled C&C traffic

Markus Leira Asprusten[1], Julie Lidahl Gjerstad[1,2], Gudmund Grov[1,2], Espen Hammer Kjellstadli[1], Robert Flood[3], Henry Clausen[3], and David Aspinall[3,4]

[1] Norwegian Defence Research Establishment (FFI), Kjeller, Norway,
{Markus.Asprusten,Gudmund.Grov,Espen-Hammer.Kjellstadli}@ffi.no
[2] University of Oslo, Oslo, Norway, julielgj@ifi.uio.no
[3] University of Edinburgh, Edinburgh, United Kingdom,
{s1784464,Henry.Clausen,David.Aspinall}@ed.ac.uk
[4] The Alan Turing Institute, United Kingdom

Abstract. A challenge for data-driven methods for intrusion detection is the availability of high quality and realistic data, with ground truth at suitable level of granularity to train machine learning models. Here, we explore a container-based approach for simulating and labelling C&C traffic of real malware through a proof-of-concept implementation.

1 Introduction & motivation

Data-driven methods for intrusion detection rely on high quality and realistic data in which to infer their detection models from. A SOC or CERT will typically have access to large quantities of log or network data which could be utilised for unsupervised or self-supervised learning. However, evaluation of such models is challenging without an evaluation set with ground truth, while supervised methods require that all data points used for learning are labelled.

One way to achieve such labels is active learning, where a security analyst is in-the-loop during learning. Another option is to label existing data, e.g. from historic incidents or labelling data points from scratch. Due to the shear size of the data, and skills required to do so, manual labelling from scratch is not feasible and would instead require automated labelling techniques, such as Snorkel [6], which provides lower quality *weak labels*. In this paper, we focus on a third approach, which is to simulate benign and adversarial behaviour, and use the knowledge from setting up the simulation to label the data. We limit the work to network-based intrusion detection systems (NIDS), and detection of command and control (C&C) beaconing traffic – a detection problem where NIDS are known to be applicable [4].

An important property of applied machine learning is the ability to generalise beyond the training data, and a known problem for NIDS is that good results on training sets often does not generalise well to an operational setting. Existing tools for C&C, often developed for penetration testing, are often used in real malware [7]. We therefore utilise such existing C&C tools in our simulation. In addition to provide realistic traffic, it means that even if a machine learning model does not generalise well beyond the tools used for simulation, it may still be valuable operationally as the tool may be used in real malware. There are several such tools: Simuland[5] focuses on Microsoft Defence products, with a mapping to the Mitre ATT&CK Framework[6]; Cobalt Strike[7] provides a number of red team activities, from generating phishing emails to browser pivoting; Metasploit[8] supports the full attack scenario, from scanning for vulnerabilities to collecting credentials and generating a final report of the attack; PoshC2[9] focuses on post-exploitation and lateral movements with encrypted C&C traffic and features extensive logging of every action and response; Covenant[10] is a .NET C&C framework; Sliver[11] is a C&C red teaming tool; Atomic Red Team[12] is a library of simple detection tests mapped to the MITRE ATT&CK framework; and Merlin[13] is a popular post-exploitation C&C Tool.

None of these tools provide ground truth and a second challenge is correct labelling captured data with suitable granularity. One common approach, used e.g. by Garcia et al [3], is to label all data from malware infected machines as 'malicious' and everything else as either 'benign' or 'background'. This will entail that some benign data is erroneous labelled as malicious. A different approach is taken by Landauer et al [5], which combines knowledge of attack time with domain knowledge of the attack steps to label post-simulation – a similar approach is also taken by Buchanan et al [1]. Their approach does not target NIDS, and in addition, the labelling quality will depend on the domain knowledge and how it is implemented in the labelling process.

In our work, we instead build on the *DetGen*-tool by Clausen et al [2], where we can achieve finer grain control of the labelling compared with Garcia et al without the need for encoding domain expertise of each attack steps. Here, we encapsulate the malware in a container and label at the container-level, thus separating traffic arising from the malware from traffic arising from other processes in a machine. We extend [2] by encapsulating the Merlin C&C simulation tool in the container. Note that whilst Cobalt Strike was the most common tool used in malware in Recorded Future's 2021 report [7], it requires a licence. We therefore use Merlin, which is also a popular tool for simulating C&C.

2 An experiment using DetGen[Merlin] with Ghost

The Merlin C&C-framework has two main components: a server and a client. The server is configured to listen for HTTP-connections from the client, and sends C&C-commands to the client over this connection. The client software runs post-exploitation on a system you wish to control and will repeatedly connect to the server with a certain interval, also called a heartbeat. To avoid detection, the interval can be skewed to vary the interval.

The DetGen framework is built around Docker-Compose[14]. Each component in DetGen runs in a container, with a separate associated container to capture

[5] https://github.com/Azure/SimuLand

[6] https://attack.mitre.org/

[7] https://www.cobaltstrike.com

[8] https://www.metasploit.com/

[9] https://poshc2.readthedocs.io/en/latest/

[10] https://github.com/cobbr/Covenant

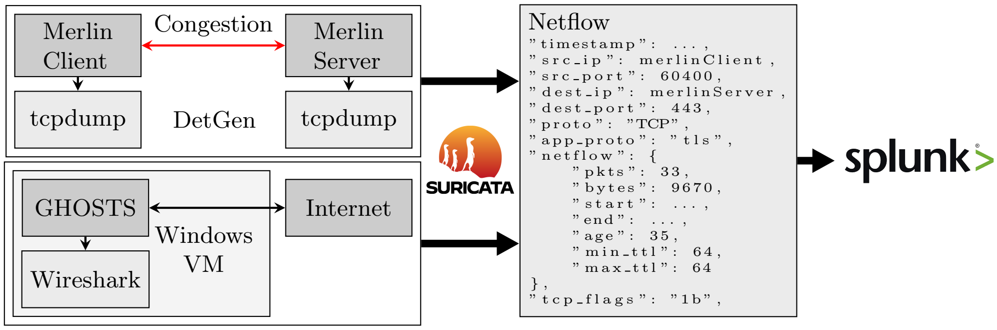
[11] https://github.com/BishopFox/sliver/

[12] https://github.com/redcanaryco/atomic-red-team

[13] https://github.com/Ne0nd0g/merlin

[14] https://docs.docker.com/compose/

the network traffic using `tcpdump`. This separation into container is then utilised when labelling traffic. In the experiment, the Merlin client and server ran in separate containers, with their associated ”`tcpdump`-containers” connected directly to the network interface of the Merlin containers In addition, the DetGen framework adds congestion and other small errors to make the simulation more realistic. This is illustrated in figure 1 (top-left).



```
"timestamp": ...,
"src_ip": merlinClient,
"src_port": 60400,
"dest_ip": merlinServer,
"dest_port": 443,
"proto": "TCP",
"app_proto": "tls",
"netflow": {
    "pkts": 33,
    "bytes": 9670,
    "start": ...,
    "end": ...,
    "age": 35,
    "min_ttl": 64,
    "max_ttl": 64
},
"tcp_flags": "1b",
```

Fig. 1. Experimental setup of Merlin with DetGen, and GHOSTS

The setup is sufficient to train a ”signature-model” which can recognise C&C-traffic. However, if our aim is to use supervised learning to train a classifier, we also need benign traffic. To achieve this, we used a framework called General Hosts (GHOSTS) [8] to simulate benign traffic. We configured GHOSTS to visit a list of domains known to be benign and record the traffic (using Wireshark), meaning, as with Merlin, we are capturing HTTP-traffic. Due to technical reasons and time constraints, we did not integrate GHOSTS into the DetGen framework, and instead captured GHOSTS traffic separately in a Windows virtual machine. This is sufficient for our proof-of-concept but will need to be integrated in the future. GHOSTS was then used to to connect to live domains with real world congestion applied. Figure 1 shows the full experimental setup for Merlin and GHOSTS. After running the simulation we changed the IP address of the Merlin Client to be the same as the GHOSTS Client. We then used Suricata[15] to convert the data sets to Netflow before combining them and importing them to Splunk[16] for visualisation. Up to this point, the C&C and Merlin traffic were in separate files, which we exploited when labelling during import into Splunk.

Splunk can then be used to train classification models, which can further be applied to real data. Here, we only visualise the traffic to illustrate how the labels can separate the traffic, as shown in figure 2. The plot shows both the number of bytes transmitted from the Merlin Client to the Merlin Server and benign traffic generated by GHOSTS (in logarithmic scale). It is easy to see the heartbeat in the graph from the Merlin Client. Note that there is some discrepancies initially

[15] https://suricata.io/

[16] https://www.splunk.com/

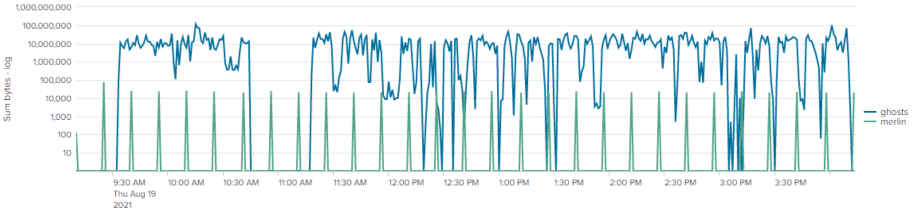


Fig. 2. #bytes of Merlin C&C traffic compared to traffic from GHOSTS in Splunk.

in the heartbeat due to some issues during setup, while a sudden dip in GHOSTS is likely caused by a need for user input to a CAPTCHA or similar.

3 Discussion and further work

We have shown that C&C tools used in practice can be simulated and labelled in a way that it can be separated from benign traffic in a SIEM with a fine grain of atomicity, which can further be utilised to train machine learning models for NIDS. Whilst the scientific contribution presented here may be limited, we believe our approach is promising for applying underlying research in an operational setting. This will require that the simulations are ran over longer time periods, using different C&C tools, different configuration and different architectures. This is also the case for the simulated benign traffic, where GHOSTS need to be integrated into the DetGen framework.

References

1. Buchanan, M., Collyer, J.W., Davidson, J.W., Dey, S., Gardner, M., Hiser, J.D., Lang, J., Nottingham, A., Oprea, A.: On generating and labeling network traffic with realistic, self-propagating malware. arXiv preprint arXiv:2104.10034 (2021)
2. Clausen, H., Flood, R., Aspinall, D.: Traffic generation using containerization for machine learning. arXiv preprint arXiv:2011.06350 (2020)
3. Garcia, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. computers & security **45**, 100–123 (2014)
4. Hutchins, E.M., Cloppert, M.J., Amin, R.M., et al.: Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research **1**(1), 80 (2011)
5. Landauer, M., Skopik, F., Wurzenberger, M., Hotwagner, W., Rauber, A.: Have it your way: Generating customized log datasets with a model-driven simulation testbed. IEEE Transactions on Reliability **70**(1), 402–415 (2020)
6. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. In: Proc. of the VLDB Endowment. vol. 11, p. 269. NIH Public Access (2017)
7. Recorded Future by Insikt Group: Adversary infrastructure report 2020: A defender's view. Tech. Rep. CTA-2021-0107 (2021)
8. Updyke, D., Dobson, G., Podnar, T., Earl, B., Cerini, A.: Ghosts in the machine: A framework for cyber-warfare exercise npc simulation. Tech. Rep. CMU/SEI-2018-TR-005, SEI/CMU (December 2018)

Elektronisk personellkontroll-system for Marinen

Truls Fismen[1], Andreas Reiming[1] og Kirsi Helkala[1][0000-0003-3698-4585]

[1] Forsvarets Høgskole, Cyberingeniørskolen, Lillehammer, Norge
truls.fismen@gmail.com, andreas.reiming@gmail.com,
khelkala@mil.no

Abstrakt. Marinens fartøy opererer med ulike beredskapsnivåer i sjøen. Ved høyeste beredskap, såkalt «klart skip», eller ved brann og/eller havari er det tidskritisk at fartøyspersonellet mønstrer på sine predefinerte plasser og at fartøysledelsen raskt kan gjøre opp personellstatus. I noen tilfeller opplever fartøysbesetningene problemer med å oppnå personellkontroll på en effektiv måte. I dag er dette en prosess som tar lang tid ved at man manuelt må kontrollere personell på utestasjoner, for så å rapportere dette inn til operasjonsrom ved hjelp av telesamband. Denne artikkelen presenteres utviklingsprosjektet av et elektronisk system for personellkontroll, som er utviklet for å tilfredsstille Marinens brukerkrav og tekniske krav, og nasjonale juridiske krav for systemer som behandler personopplysninger. Funnene er basert på Skjold-klasse korvett, men kan generaliseres til samtlige av Marinens seilende plattformer.

Nøkkelord: Personellkontroll, personvern, informasjonssikkerhet.

1 Introduksjon

Marinen opplever problemer med å oppnå personellkontroll om bord på sine fartøysklasser på en effektiv måte. I dag er dette en prosess som tar lang tid ved at man manuelt må finne status på personell. Motivasjonen for prosjektet kommer fra viktigheten av å kunne opprette personellkontroll på kortest mulig tid om bord på Marinens fartøy under tidskritiske og uoversiktlige forhold. Et konkret eksempel på en situasjon hvor et slikt system kunne ha hatt verdi er Helge Ingstad-ulykken i 2018. Her opplevde skipssjefen at det tok lang tid å oppnå kontroll på personellet under en svært stressende og uoversiktlig situasjon. Tidligere skipssjef Preben Ottesen på KNM Helge Ingstad uttalte til VG i etterkant av hendelsen at det i fredstid er førsteprioritet å få kontroll på mannskapet under ulykker. Dette tar vanligvis noen få minutter, men under havariet opplevde han at dette tok mye lenger tid enn vanlig [23]. Påvirkende årsak var blant annet strømbrudd og utfall av fartøyets sambandssystemer og dermed manglende evne til å rapportere status, i tillegg til at flere besetningsmedlemmer var innesperret på lugarer og følgelig ikke hadde mulighet til å melde fra.

Denne artikkelen er forkortelse av en utviklingsoppgave som ble gjort som en bacheloroppgave ved Cyberingeniørskolen [15]. Opprinnelig oppgave har ikke blitt offentlig publisert fordi den inneholder begrenset informasjon. Informasjon som er gradert eller unntatt offentlighet er utelatt og ikke å bli diskutert i denne artikkelen.

Artikkelen presenterer besvarelsen til oppgavens problemstilling: «Hvilke krav stilles til et elektronisk system som skal gi informasjon om personellstatus i form av personelltilstand og posisjon, og hvordan kan et slikt system utvikles og implementeres på korvetter i Skjold-klassen?»

2 Bakgrunn

Selv om prototyputvikling er en teknisk prosess, må de tekniske løsningene for et elektronisk system som behandler om personopplysninger tilfredsstille juridiske krav. I resterende del av kapitlet gjennomgås de juridiske aspekter og lover som er sentrale for valgene gjort under utviklingsprosessen. I tillegg må informasjonssikkerhet ivaretas når systemer blir utviklet [10, 17].

2.1 Personvernforordringen

Personvernforordningen er en forordning for å regulere behandling av personvernopplysninger i EØS. Dette er i norsk lov gjennomført gjennom personopplysningsloven av 2018. I forordningen blir begrepet «behandling» definert til blant annet å omfatte innsamling, registrering, lagring og bruk av personopplysninger. Begrepet «personopplysninger» omfatter opplysninger om en identifisert eller identifiserbar fysisk person, jf. art. 4 nr. 1. Et sentralt aspekt ved forordningen og personvernopplysningsloven er at organisasjoner, eksempelvis Forsvaret, må inneha et legitimt behandlingsgrunnlag for å kunne behandle personvernopplysninger, jf. art. 6 nr. 1.[21]

Behandlingsgrunnlagene for å behandle personvernopplysningen finner man i art. 6 nr. 1 bokstav a) til f) i personopplysningsloven [21]. Et sentralt poeng er at det kan finnes flere behandlingsgrunnlag som kan passe til et enkelt formål, men man kan kun ha et behandlingsgrunnlag per formål. Under er en liste over de ulike behandlingsgrunnlagene med en beskrivelse hentet fra art. 6 nr. 1 a) til f) i personopplysningsloven.

a) Den registrerte har samtykket til behandling av sine personopplysninger for ett eller flere spesifikke formål.
b) Behandlingen er nødvendig for å oppfylle en avtale som den registrerte er part i, eller for å gjennomføre tiltak på den registrertes anmodning før en avtaleinngåelse.
c) Behandlingen er nødvendig for å oppfylle en rettslig forpliktelse som påhviler den behandlingsansvarlige.
d) Behandlingen er nødvendig for å verne den registrertes eller en annen fysisk persons vitale interesser.
e) Behandlingen er nødvendig for å utføre en oppgave i allmennhetens interesse eller utøve offentlig myndighet som den behandlingsansvarlige er pålagt.
f) Behandlingen er nødvendig for formål knyttet til de berettigede interessene som forfølges av den behandlingsansvarlige eller en tredjepart, med mindre den registrertes interesser eller grunnleggende rettigheter og friheter går foran og krever vern av personopplysninger, særlig dersom den registrerte er et barn.

Det er verdt å nevne at ved bruk av art.6 nr.1 bokstav c) eller e) som behandlingsgrunnlag, må det også finne hjemmel i norsk lov [21], f.eks. sikkerhetsloven [20].

2.2 Grunnleggende personvernprinsipper

Datatilsynet har formulert en følgende personvernsprinsipper for å ivareta lovlig, rettferdig og transparent behandling av personvernsopplysninger [9]. Disse er basert på artikkel 5, 6 og 9 i personvernforordningen.

Formålsbegrensning: Formålet med behandlingen av personvernsopplysningene må være identifisert og beskrevet presist. Dette må gjøres på en entydig måte slik at alle interesserte har lik forståelse for hva disse opplysningene skal brukes til. Dette betyr også at dataene ikke kan brukes til andre formål enn de som er beskrevet med mindre det hentes inn nytt samtykke.

Dataminimering: Prinsippet om dataminimering innebærer å begrense innsamlet data til det minimum som er nødvendig for formålet. Dette betyr at identitetsopplysninger som ikke er relevant for formålet ikke skal innsamles. Det kan også være relevant å begrense hvor mange og hvilke personer det samles inn opplysninger om.

Riktighet: Personopplysninger som behandles skal være korrekte og nøyaktige. Hvis det er nødvendig, skal de oppdateres. Dette innebærer at den med behandlingsansvar har ansvar for å slette og endre uriktige personopplysninger, også i backuper.

Lagringsbegrensning: Prinsippet om lagringsbegrensing betyr at informasjon bare skal lagres så lenge som nødvendig og deretter slettes eller anonymiseres automatisk.

Integritet og konfidensialitet: Dette prinsippet innebærer at integriteten og konfidensialiteten bevares. Dette betyr at det skal være iverksatt tiltak for at utenforstående ikke får tilgang til å lese eller endre informasjonen som er lagret.

Ansvarlighet: Ansvarlighet betyr at den ansvarlige må opptrå i samsvar med gjeldene regler for behandling av personopplysninger. Det er ikke nok å ha ansvaret, man må også vise at man tar det alvorlig. Dette innebærer å etablere tekniske og organisatoriske tiltak for å sørge for at opplysningene blir ivaretatt på en ansvarlig måte.

2.3 Skjold-klasse

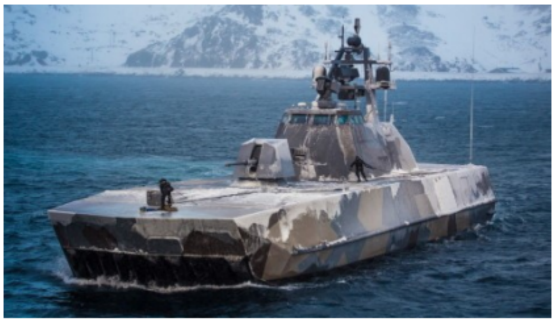


Fig. 1. Skjold-klassen korvett [6]/Henriette Dæhli/Forsvaret.

Denne artikkelen er basert på Marinens Skjold-klasse fartøy, illustrert på fig. 1. Fartøyene er en blanding mellom luftputebåt og katamaran, noe som muliggjør at man kan løfte fartøyet med vifter under skrogkonstruksjonen. Dette minsker fartøyets motstand i vannet og muliggjør høy hastighet. De er bygget med en «stealth teknologi» som gjør at de er vanskelig å oppdage på radar og det blir benyttet et lett komposittmateriell for å oppnå lav vekt. Skjold-klassen er noen av verdens raskeste militære fartøyer og kan i hovedsak bekjempe overflatetrusler, men innehar noe anti-luftkapasitet. Selv om artikkelen fokuserer om Skjold-klassen, kan ideen generaliseres til andre fartøy. Selv implementasjonen må likevel testes, fordi material av fartøy kan påvirke signaler.

3 Metode

Prosjektet har vært et utviklingsprosjekt og benyttet en iterativ og smidig metodikk med fire faser. I kravinnhentingsfasen ble det utarbeidet hvilke krav som stilles til et elektronisk system for personellkontroll. Det ble kartlagt bruker- og tekniske krav til systemet via samtaler med teknisk personell om bord. Intervjugodkjenning ble gitt av Norsk senter for forskningsdata og Forsvarets forskningsnemd. Videre utledet vi juridiske krav som stilles til et slikt system under veiledning av jurist i Cyberforsvaret. Kravene blir presentert i kap. 4.

Informasjonsinnhentingsfasen handlet om sambandsmidlene som eksisterer på korvetter i Skjold-klassen (gradert) og muligheten for å utvikle et eget system. Kompetansehevingen ble i denne fasen gjennomført ved litteraturstudie. Inspirasjon for å utvikle en personellkontroll system ble hentet fra åpne kilder på internett som forklarte hvordan lignende systemer lages [16, 27, 29, 31]. I tillegg, har flere tekniske sider blir benyttet for eksempel [4, 5, 12, 18, 25, 26, 28].

I utvikling og implementasjonsfasen ble det utviklet en prototype av et elektronisk system for personellkontroll til bruk på Skjold-klassen ved hjelp av smidig systemutviklingsmetodikk. Fordelen ved å benytte denne metoden mot den mer tradisjonelle systemutviklingsmetoden er i hovedsak at man kan ta høyde for oppdukkende krav og behov som stilles til systemet [29]. Ulempen med denne metoden kan være at man styrer mot kundenes subjektive krav, og at disse ikke nødvendigvis realiserer målene som stilles til et system. Prototypen blir presentert i kap. 5.

Til slutt ble prototypen til systemet testet og evaluert. Testene fokuserte på hvordan systemet oppfyller kravene som er utledet fra de tidligere fasene. Testresultatene er presentert i kapittel 6, mens evaluering av prototypen er presentert i kap. 7. Her diskuter vi systemet oppimot de ulike kravene som ble utledet i fase én. I tillegg til dette diskuterer vi hvilke sikkerhetstrusler som vil være relevante for et slikt system. Opprinnelig var intensjonen å utføre tester om bord på fartøyet, men dette var ikke mulig på grunn av Covid-19 situasjonen. Testene ble derfor gjennomført på en bygning.

4 Resultat I: Krav

Kapitlet presenterer resultatene i følgende rekkefølge: Først tar vi for oss de ulike kravene utledet til det elektroniske systemet for personellkontroll både bruker-, tekniske- og juridiske krav. Etter dette presenteres utviklet elektronisk system for personellkontroll. Argumentasjon for valgene som har blitt tatt kommer etter denne. Til slutt legges det frem resultater fra testene som ble gjennomført.

4.1 Kravspesifisering

Basert på samtaler med teknisk personell og egne erfaringer ble det utledet bruker krav og tekniske krav. Kravene er selvforklarende og blir ikke diskutert i detalj her.

1. Systemet skal kunne gi posisjon og status på personellet.
2. Posisjon bør være nøyaktig ned til hvilket rom personen befinner seg, i hvert fall til en sone.
3. Det skal være enkelt å benytte systemet, og det skal ikke komme i veien.
4. Batteritid, i tilfell strøm faller ut, for alle leddene i systemet må være så god at mannskapet ikke må tenke på dette.
5. Systemet må ikke kunne oppdages av andre fartøy.
6. Systemet skal ikke interferere med andre systemer om bord på fartøyet.

4.2 Personvern

Formålsbegrensning setter krav om at formålet med innhentingen av personopplysningene må være strengt definert og informasjonen vi lagrer så lite som det er behov for. Det er også kritisk at dataensom blir innhentet ikke skal benyttes til andre formål enn det som er beskrevet. Formålet med det elektroniske systemet er å sørge for personellkontroll ved stressende og kritiske situasjoner som de ulike beredskapstilstandene ved eksempelvis «klart skip», «brann» og «havari».

1. Systemet må ha en funksjon som gir mulighet for å skru det av og på.
2. Dataen som systemet innsamler skal kun benyttes til å opprette personellkontroll ved situasjoner hvor dette er viktig for liv og helse om bord, også under trening.
3. Alle på fartøyet skal bli informert når systemet aktiveres.
4. Posisjon og statusdata skal ikke lagres av systemet.
5. Mengden personopplysninger må minimeres slik at man kun henter inn informasjon som er kritisk for at systemet skal levere tilstrekkelig funksjonalitet.

Årsaken for første kravet er at systemet kun skal benyttes i situasjoner hvor det er kritisk å opprette personellkontroll, og det derfor ikke vil være nødvendig at det er på til alle tider. For å følge dette prinsippet bør systemet inneha denne funksjonen.

Det andre kravet stilles fordi det understøtter prinsippet om at man kun benytter dataen til formålet. Det kan tenkes at uten et krav som dette vil løsningen kunne misbrukes. Eksempelvis at systemet blir benyttet for å overvåke personellet, og vurdere effektiviteten i arbeidet deres [30]. Dette må man unngå ved å sette strenge krav til formålet og når systemet skal kunne tas i bruk. I dette tilfellet er ikke systemet ment for å fungere som en effektivitets måler og man må av den grunn minimere sjansen for at det kan misbrukes slik.

Det tredje kravet kan praksisers enten at det sies ifra på forhånd i hvilke situasjoner systemet tenkes å nyttes, eller eksempelvis at man melder at det tas i bruk over høyttaler også kjent som PA-anlegg. Dette er igjen på bakgrunn av at personellet om bord på fartøyet skal være klar over når de blir overvåket.

Lagringsbegrensning setter krav om at informasjonen på systemet kun lagres så lenge det er nødvendig og at det deretter slettes eller anonymiseres [9]. Likevel kan man argumentere for at det ville det kunne være nyttig å lagre data en kort periode, f.eks. benytte den i etterforskning ved ulykker eller lignende. Dette kunne blitt gjort ved å skrive ned en periode man lagrer dataen i en samtykkeerklæring.

Det femte kravet baseres på dataminimering prinsippet som sier at datainnsamlingen og personopplysninger må begrenses til kun det som er nødvendig for formålet [9]. Det er verdt å nevne at dette prinsippet understøtter krav tre, men også blir understøttet av prinsippet om formålsbegrensning.

4.3 Behandlingsgrunnlag

Bruken av systemet vil medføre behandling av personopplysninger. Av denne grunn må det bli vurdert om Forsvaret har såkalt behandlingsgrunnlag etter art. 6 nr. 1 i Personvernforordningen (se kap. 2.1). I art. 6 nr. 1 a) til f) er det syv alternativer man kan benytte for å argumentere for behandlingsgrunnlag, og man kan kun benytte seg av én av disse. Vi har vurdert grunnlag a) og d) som sentrale for dette systemet.

Det kan argumenteres for at grunnlag a) kan benyttes da denne omhandler at en virksomhet kan behandle personopplysninger om de har innhentet samtykke fra de det gjelder. Det er sentralt at samtykke er frivillig og at vedkommende er godt informert, samt at formålet er godt spesifisert [8]. Dette vil kunne være mulig å få til ved å lage et skriv om hvilke personopplysninger som skal innhentes, formålet for innhentingen av informasjonen og hva det skal brukes til. Likevel kan det være problematisk at personell ikke egentlig får en reel mulighet til å gi samtykke, f.eks. hvis samtykke må gis for å kunne jobbe på fartøyet, blir det vanskelig å bedømme hvorvidt samtykket er frivillig.

Grunnlag d) kan benyttes som behandlingsgrunnlag. Dette behandlingsgrunnlaget går ut på at det må være nødvendig for å «verne den registrertes eller en annen fysisk persons vitale interesser» [8]. Denne er veldig aktuell å benytte for det elektroniske systemet på bakgrunn av at formålet med systemet er knyttet til liv og død, altså i hovedsak fare for helsen. Dette behandlingsgrunnlaget blir av Datatilsynet beskrevet som svært snevert [8]. Likevel mener vi at liv og helse på Marinens fartøy, spesielt i kritiske situasjoner som klart skip, havari og brann havner innenfor denne tolkningen, og dermed er grunnlag d) det beste behandlingsgrunnlaget for det elektroniske personellkontroll systemet.

5 Resultat II: Elektronisk system for personellkontroll

Basert på kravene vist i kap. 4 har vi laget en prototype på et innendørs posisjoneringssystem med funksjoner for personellkontroll. Det overordnede målet til systemet er å kunne gi informasjon om hvilket rom eller sone personellet på fartøyet befinner seg i, og en mulighet til å enkelt kunne melde ifra om egen tilstand.

Overordnet er systemet bygd opp slik at hver person utrustes med en liten BLE-beacon (Bluetooth Low Energy Identifier) [26] som bæres på kropp, for eksempel rundt halsen. Fartøyet blir utrustet med en sender og mottaker i hvert rom eller i en definert sone. Denne sender og mottakeren består av maskinvare og programvare som skanner etter BLE-beacons, henter ut informasjon og sender denne informasjonen videre til en sentralisert hub via rutere utstasjonert i fartøyet. Informasjonen den sender videre er informasjon om RSSI-verdien (Received Signal Strength Indicator) [22] og BLE-beaconets UUID (universally unique identifier) [26], samt rommets ID. Denne sentraliserte huben prosesserer dataen og fremviser posisjonen til de ulike personene og tilstand på en skjerm som brukeren kan nyttiggjøre seg av. En overordnet oversikt over systemet er vist på fig. 2.

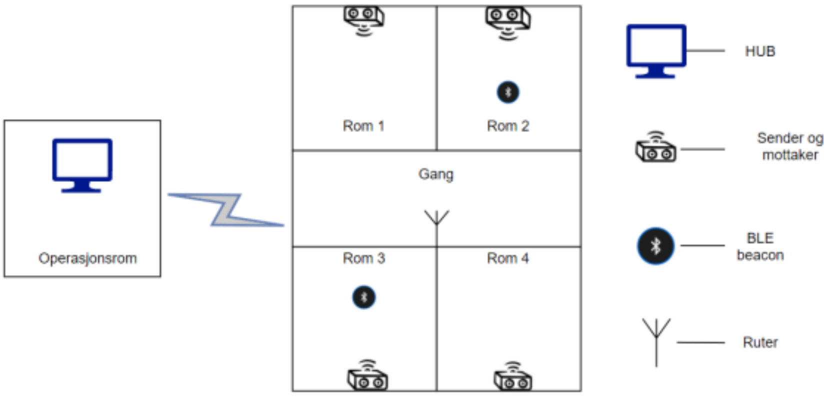


Fig. 2. Overordnet oversikt over systemet.

Prototypen består av fem sensorsystem er med tilhørende XBee-moduler (se fig. 3) og en hub med en XBee-modul koblet til en Arduino UNO (se fig. 4), som igjen er tilkoblet en Raspberry Pi. Størrelsen på sensorsystemene er omtrent 10 cm x 15 cm. Størrelsen på BLE-beacon (se fig. 5) er omtrent 3cm x 3 cm.

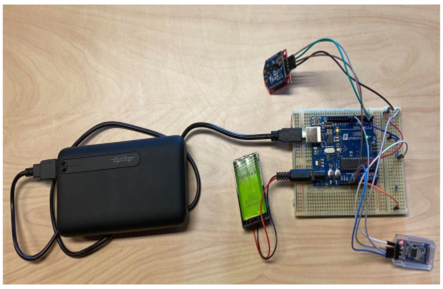


Fig. 3. Sensorsystemet.

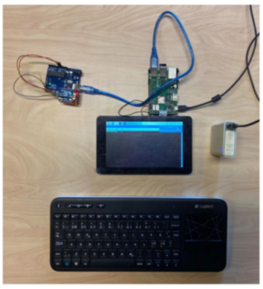


Fig. 4. Hub.

Fig. 5. BLE.

5.1 Sensorsystemet

Maskinvaren som blir benyttet i sensorsystemet er HM-10 moduler [2], Arduino UNO [5, 14] og BLE-beacons [1]. I tillegg består sensorsystemet av en strømforsyning som er tiltenkt å koble i fartøyets strømnett, og en batteripakke med et 9 volt batteri som kan overta med en gang strømmen om bord faller ut.

HM-10 benyttes for å skanne etter BLE-beacons. For å hente ut dataen fra HM-10 må man laste inn en programkode på Arduino UNO [4]. Denne koden henter først informasjonen fra HM-10 modulen ved å sende kommandoen «AT+DISI?» som returnerer alle Bluetooth-enheter og gir informasjon om disse på et ibeacon format [3]. Videre sorterer koden ut UUID og RSSI verdien til oppdagede BLE-beacons. I tillegg legger koden til IDen til rommet HM-10 modulen befinner seg. Dette for å kunne knytte rom oppimot RSSI-verdi i huben. Denne rom-IDen er et tall på to bytes som er unikt for hvert rom og legges inn når arduinoene programmeres. Informasjonen koden sorterer ut lagres i en variabel slik at den kan hentes ut for å videresendes til huben.

Nettverksstrukturen til sensorsystemet er et stjernenettverk, fordi de forskjellige BLE-beacons kommuniserer med HM-10 modulen uten å kommunisere med hverandre [16, 18]. HM-10 modulen vil være koordinatoren i nettet mens BLE-beacons vil være ulike noder i nettverket som kommuniserer med koordinatoren.

5.2 Transmisjon

Maskinvaren som blir benyttet i transmisjonsdelen er XBee-moduler [11, 13]. Disse brukes for å sende informasjonene fra sensorsystemet videre til den sentraliserte huben. Disse er koblet til de samme Arduino UNO som benyttes i sensorsystemet. Til prototypen er det anskaffet XBee PRO S2C-moduler og adaptere. Disse adapterene benyttes for å sørge for enkel og stabil kommunikasjon mellom modulene og Arduino UNO.

XBee-modulene konfigureres via det tilhørende programmet XCTU. Det lastes inn en fastvare på modulene for at de kan operere i et maske-nettverk og videre det konfigureres innstillinger for å tillate kommunikasjon mellom modulene og arduinoene.

En av modulene blir konfigurert til koordinator i nettet, mens resten blir konfigurert til rutere. Programvaren for å sende dataen fra sensorsystemet som benyttes i transmisjonsdelen er sammenflettet med programkoden på Arduino UNO som benyttes for å hente ut informasjonen fra sensorsystemet. Her benytter vi et eget bibliotek på programmeringsmiljøet til arduinoen, Arduino Integrated Development Environment (IDE) [24] for å lage en kode som fyller datapakker med informasjonen som hentes ut i sensorsystemet og videresender dette til huben.

XBee modulene som blir benyttet i denne prototypen er konfigurert til å operere i et maske-nettverk. Den ene XBee modulen i nettverket som er konfigurert til å være koordinator vil motta alt av datapakker som enten sendes fra et sensorsystem innen rekkevidde eller som er videresendt av ulike sensorsystemer eller egne rutere.

XBee-modulene koblet til sensorsystemet vil være konfigurert til å fungere som rutere. Dette vil si at de sender ut dataen de får fra det tilkoblede sensorsystemet, i tillegg til å videresende data de mottar fra andre XBee-moduler tilknyttet andre sensorsystemer. Det er også mulig å konfigurere XBee-moduler til ruter uten tilknytning til et sensorsystem. Denne vil kun videresende datapakker den mottar og er tiltenkt å fungere som reelle hvis man har dårlig forbindelse i noen områder.

5.3 Hub

For at en bruker av prototypen skal kunne nyttiggjøre seg av informasjonen som hentes har vi laget en hub som kan presentere dataen som innhentes fra forskjellige områder i fartøyet. Det er definert tre operasjoner som må bli utført i huben. Først må dataen koordinator XBee-modulen mottar hentes ut via en tilkoblet Arduino UNO. Deretter må denne sensordataen analyseres av et program for å finne posisjon og tilstand til personellet. Etter dette må informasjonen fremvises, slik at det den kan benyttes for å forbedre personellkontroll.

En XBee-modul som er konfigurert til koordinator benyttes for å motta all informasjon som sendes i nettverket. Denne XBee-modulen er koblet til en Arduino UNO. Arduinoen har et program som leser dataen mottatt fra XBee-modulen og sender den ut på USB-porten som serielldata.

Serielldataen mottas av Raspberry Pien [25] og analyseres ved hjelp av Python-kode. Raspberry Pien er koblet til en tilhørende 8 tommers skjerm. Til prototypen er det anskaffet én Raspberry Pi [28] og en tilhørende 8 tommer skjerm.

På Raspberry Pien settes det opp kommunikasjon til arduinoen for å motta meldinger via UART-protokollen. Meldingene sjekkes for om de er på riktig format ved en regulærtuttrykksjekk. Hvis formatet er riktig, sorteres meldingen basert på rom-ID (to første bytes) og UUID (neste 20 bytes). Finnes det en oppføring for denne UUIDen og rom-IDen oppdateres RSSI verdien, ellers legges det til en ny verdi. Samtidig lagres det tidspunktet verdien ble lagt til. Neste gang det kommer inn en melding som gjelder samme UUID vil programmet sjekke om det har verdier som er eldre enn ett minutt og slette disse. Dette er for å unngå gamle utdaterte verdier i systemet. Til slutt sorterer programmet hvilket rom som har best signalstyrke til beaconet basert på RSSI og skriver ut hvilket rom beaconet befinner seg.

5.4 Argumentasjoner for metoder, protokoll og maskinvare brukt i prototypen

Det ble valgt å benytte signalmåling basert på signalstyrke over metoder basert på tid. En av de grunnene var at det finnes mye kommersiell hyllevare man kan benytte for å finne informasjon om signalstyrke. En annen grunn var at metoder basert på tid krever mer avanserte komponenter, noe som kan føre til at systemet blir dyrere og større. Dette vil kunne gå imot kravet om at systemet ikke skal være i veien for personellet. Ulempen med signalmåling basert signalstyrke er at systemet kan bli noe mer unøyaktig.

Det ble valgt å benytte nærhet posisjoneringskalkulering over trilaterasjon og triangulering. Den mest definerende grunnen er at BLE-beacon som benyttes i systemet ikke trenger å ha forbindelse med flere sensorsystemer samtidig, og det er nyttig i et fartøysmiljø. En annen grunn er at den ikke trenger alt for komplisert programvare. Ulempen med denne metoden er at den gir noe mer unøyaktig posisjon enn de andre metodene.

Prototypen til systemet er bygget opp av stjerne -og maskenettverktopologi. Stjernetopologi er valgt til sensorsystemet fordi dette muliggjør at man kan hente ut informasjon om noder i nettverket fra en koordinator. I tillegg er det mye kommersiell maskinvare som støtter denne topologien og lite konfigurasjon som kreves for å realisere denne topologien. Masketopologi er valgt til sensorsystemet fordi dette lager et redundant nettverk. Informasjonen fra sensorsystemene vil kunne sendes over maskenettverket. Rekkevidden til systemet vil også øke på bakgrunn av at man kan sende informasjon via noder til destinasjonen. I likhet med stjernetopologi finnes det mye kommersiell maskinvare som støtter denne topologien.

Protokollene som er benyttet i prototypen er BLE [19] og Zigbee [12]. BLE benyttes til sensorsystemet og Zigbee til transmisjon fra sensorsystemene til huben. Overordnet er det ingen av disse protokollene som interferer med andre systemer på Skjold-klassen, noe som er et viktig argument for at de er valgt til systemet.

BLE er valgt til sensorsystemet fordi den støtter stjernetopologi og det finnes mye kommersiell maskinvaresom benytter seg av denne protokollen. I tillegg er UUID-formatet man identifiserer noder med veldig verdifullt for å kunne identifisere personell, og for å utvikle programvare. Videre er protokollen utviklet for lavt energiforbruk, dette gjør at maskinvare protokollen benyttes på vil ha god batteritid. Rekkevidden er på 15 m innendørs, noe som er godt nok for at den skal kunne nyttes i sensorsystemet.

Zigbee er valgt til transmisjonsdelen fordi, for det første, protokollen støtter maske-topologi og teoretisk opp til 64000 noder i et nettverk [12] og den har en teoretisk rekkevidde på 30 m innendørs ved 2400 MHz versjonen. Et annet element er at det finnes mye kommersiell maskinvare man kan benytte. Det er verdt å nevne at ved 2400 MHz kan den interferere med Wi-Fi. Likevel skrus gjerne trådløse nettverk av under seilas, noe som kan være et argument for at denne frekvensen er gunstig å benytte.

Maskinvaren (BLE-beacons, HM-10 moduler, Arduino UNO, XBee-moduler og Raspberry Pi) er overordnet valgt fordi de støtter posisjoneringsmetoden, nettverkstopologi og protokoller. Fordel for BLE-beacon er liten i størrelse med god batteritid, og har en knapp som vi kunne programmere med ønsket funksjon. I tillegg opererer den på en frekvens som ikke interfererer med allerede eksisterende sambandsmidler. Fordel for HM-10 modulen er at den er kompatibel med BLE-beacon og har tilkoblingsmuligheter til Arduino UNO mikrokontrolleren.

XBee-modulen ble valgt fordi det er en Zigbee-modul med god rekkevidde og har gode muligheter for maskenettverk. I tillegg kan den tilkobles til Arduino UNO mikrokontrolleren og en benytter frekvenser som ikke vil påvirke de allerede eksisterende sambandsløsningene om bord, med unntak av Wi-Fi. Arduino UNO og Raspberry Pi ble valgt fordi de innehar nok datakraft for å kjøre trengte programmene.

6 Resultat III: Testing av prototypen

For å kunne vurdere prototypen ble det gjennomført tester for å evaluere systemets rettidighet og nøyaktigheten til nærhetsprogramvaren. Indirekte ble systemets rekkevidde også testet. Måling av tid ble gjort med stoppeklokke og måling av avstand med målebånd.

6.1 Test av registreringstid ved bytte av posisjon

Målet med denne testen var å finne ut hvor lang tid det tar for systemet å oppdatere ny posisjon på en BLE-beacon når den bytter rom. For å prøve å unngå feilmålinger satt vi ut to sensorsystem med god avstand på omtrent 8 meter. Startpunktet for testet er området man går ut fra sonene til det ene sensorsystemet og inn i sonen til det andre sensorsystemet. Det er her man starter stoppeklokken og måler tid fram til det oppdateres at BLE-beacon har blitt flyttet fra rom 1 til rom 2 eller motsatt. Det ble gjennomført 10 tester og resultatene er fremlagt i tab. 1. Resultatet gir gjennomsnitt 12,22 sekunder med standardavvik 4,58 sekunder.

Tab. 1. Resultater av rombytting.

Forsøk	1	2	3	4	5	6
Tid	11,14	9,79	22,63	4,67	15,91	9,45
Forsøk	7	8	9	10	$\bar{x}$	**sd**
Tid	14,02	9,02	13,83	11,73	**12,22**	**4,58**

6.2 Test av registreringstid ved innmelding av status

Ved denne testen er formålet å finne ut hvor lang tid det tar for prototypen å oppdatere innmeldingen av personellstatus fra BLE-beacon. I et forsøk på å få mest mulig nøyaktige målinger benyttet vi kun et sensorsystem og en BLE-beacon. Deretter målte vi tiden fra man trykker inn knappen på BLE-beacon til det ble fremvist på huben. Det ble gjennomført 10 tester og resultatene er fremlagt i tab. 2. Resultatet gir gjennomsnitt 11,52 sekunder med standardavvik 3,50 sekunder.

Tab. 2. Resultater av status innmelding.

Forsøk	1	2	3	4	5	6
Tid	10,23	13,34	15,63	7,19	8,76	15,76
Forsøk	7	8	9	10	$\bar{x}$	**sd**
Tid	13,43	7,23	7,67	15,92	**11,52**	**3,50**

6.3 Test av nøyaktighet til prototypen

Ved denne testen var formålet å finne ut hvor nøyaktig prototypen kan levere posisjonsdata. Testmiljøet er rom på en kaserne med størrelse på omtrent 3 meter i bredden og 6 meter i lengden. Måten testen ble gjennomført var ved at en testperson tok en BLE-beacon rundt halsen og gikk en runde i testmiljøet (se fig. 2). Personen begynte i gangen, deretter gikk personen til rom 1, rom 2, rom 3 og rom 4. Det ble også gjennomført målinger ved å legge en BLE-beacon nærmest mulig veggen mot et annet sensorsystem for å se om man får feil posisjonsdata. I tillegg til dette gjennomførte vi testen med åpne og lukkede dører.

Det overordnede resultatet fra testene er at prototypen leverer riktig posisjonsdata når man er tydelig innenfor rommet til et sensorsystem, men kan være noe unøyaktig i grensepartiet mellom flere sensorsystemer. Resultatene fra testen gjennomført med åpne dører er fremlagt på fig. 6. Områdene markert i oransje er de hvor det kan oppstå feilposisjonering, disse går omtrent en meter inn i rommet og noe utenfor.

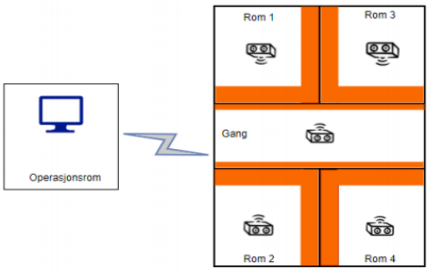


Fig. 6. Nøyaktighet ved åpne dører.

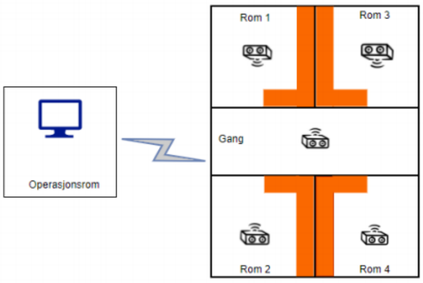


Fig. 7. Nøyaktighet ved lukkede dører.

Resultatene fra testen gjennomført med lukkede dører er fremlagt på fig. 7. Områdene markert i oransje er de hvor det kan oppstå feilposisjonering, disse går omtrent en meter inn i rommet, men her vil de avgrenses mer enn med åpne dører.

7 Diskusjon

For å vurdere prototypen skal vi ta utgangspunkt i hvordan det utfyller de ulike kravene definert for et elektronisk system for personellkontroll.

7.1 Bruker og tekniske krav

Det første bruker kravet var at systemet skulle kunne gi posisjon og status på personellet. Den utviklede prototypen vil oppfylle dette kravet ved at den gir informasjon om rom eller sone personellet befinner seg i fartøyet. I tillegg til at man kan melde inn egen status ved å trykke på en knapp på BLE-beacon. Når dette sees opp mot hvilken posisjon personell skal ha ifølge rullen til skipet, kan det bestemmes om skipet er klart eller om det kreves å sette i gang personsøk for å lokalisere vedkommende. I denne artikkelen har vi ikke testet BLE-beacon sin nyttighet i «man over bord» situasjoner.

Det andre brukerkravet var at system et burde gi posisjon som er nøyaktig nok til å levere informasjon om hvilket rom personellet befinner seg. Den utviklede prototypen oppfyller delvis dette kravet. Ut fra resultatene om nøyaktighet til prototypen ser man at prototypen kan levere unøyaktig rom data i grensepartiene mellom sensorsystemer. Av denne grunn kan man argumentere for at systemet vil fungere bedre til å levere data om sone, og ikke rom. Testene er riktignok simulert på rom med lettvegger. Av den grunn kan man argumentere for at det er vanskelig å si hvordan posisjoneringen vil være på en korvett med mulig andre vegger. Likevel ved tykkere vegger og materiale av stål vil sannsynligvis posisjoneringen bli mer nøyaktig på grunn av at strålingens diffraksjon påvirkes av materialer med høyere tetthet. Av den grunn kan det argumenteres for at nærhetsteknikken som brukes nå egner seg best på fartøy med tykke vegger, som for eksempel stålskottene på en fregatt. Ved lettvegger som eksisterer på korvett bør man mulig kunne benytte en trianguleringsalgoritme som gir bedre nøyaktighet.

Det tredje brukerkravet var at systemet skulle være lett for mannskapet å benytte, samt at det ikke skulle komme i veien. Det kan argumenteres for at prototypen oppfyller dette kravet i stor grad. Mannskapet må gå rundt med en liten BLE-beacon i et halsbånd, eller ha den liggende i uniformen. Sensorsystemene blir satt ut på strategiske steder som ikke skal føre til noen hindring og hvor det er tillat å ha sensorer. Huben plasseres i nærheten av operasjonsrommet hvor den ikke bør komme i veien for noen og er, med tanke på gradering tillat å installere, kun være et ekstra hjelpemiddel for personellkontrollansvarlig. Kravet til batteritid, tenker vi å vare tilfredsstillende i denne prototypen.

Det femte kravet er tekniske kravet som omhandlet at systemet ikke skal kunne oppdages av andre fartøy. Dette oppfyller systemet ved å benytte protokoller med lav rekkevidde. Maksrekkevidden til Zigbee ved 2400MHz, som er frekvensen vi har brukt, er på 400 meter. Dette vil være betydelig kortere enn avstanden man visuelt kan se fartøyet. Dette er allikevel ikke testet av oss med avanserte EK-sensorer, noe som bør gjøres før systemet tas i bruk.

Det siste kravet er tekniske kravet som omhandler at systemet ikke skal interferere med andre systemer om bord på fartøyet. Dette kravet oppfyller systemet ved at man har kartlagt de ulike interne sambandsmidlene om bord på fartøyet. Deretter ble det valgt protokoller som ikke interfererer med noen av de kartlagte systemene, hverken ved lik frekvens eller protokoll. Dette vil sørge for at de andre systemene om bord ikke vil bli påvirket av personellkontrollsystemet.

7.2 Juridiske krav

Det første juridiske kravet var at systemet skulle ha en funksjon som gir mulighet for å skru systemet av og på. På mange måter kan man argumentere for at prototypen til systemet oppfyller dette kravet. For å skru av huben og sensorsystemene kan man koble ut strømtilførselen. Altså har prototypen en funksjon for å skru systemet av og på, likevel vil dette være tungvint. Av den grunn kan det argumenteres for at det vil være hensiktsmessig å utvikle en mer sentralisert metode på et ferdigstilt system der man har mulighet til å skru av og på sensorene fra huben.

Det andre juridiske kravet som ble definert for systemet omhandlet at informasjonen som innhentes av systemet kun skal nyttes til å opprette personellkontroll ved situasjoner hvor dette er viktig for liv og helse, og ved trening på slike situasjoner. På skjermen til huben og det tilhørende tastaturet kan man starte og stoppe programmet som nyttiggjør seg av dataen som sendes fra de ulike sensorsystemene. Av den grunn kan man argumentere for at prototypen legger til rette for dette kravet.

Det tredje juridiske kravet omhandler at personellet på fartøyet skal bli informert når systemet aktiveres. Dette er noe som ikke direkte kan knyttes opp til systemet, men bruken av det. Ved dette kravet må det meldes over PA-anlegg at systemet startes og opplyses om at personellets posisjon overvåkes, for eksempel i begynnelse av en seilas.

Det fjerde juridiske kravet omhandler at posisjon og statusdata ikke skal lagres av systemet. Dette oppfyller systemet ved at det er utformet til å ikke lagre dataen den mottar, kun fremvise det når nødvendig.

Det femte juridiske kravet omhandler at mengden personopplysninger må begrenses til kun det som er kritisk for funksjonen til systemet. Prototypen benytter kun navn eller stilling, posisjon og tilstand. Av denne grunn kan man argumentere for at prototypen oppfyller dette kravet. I tillegg sendes hverken navn eller stilling til personen over radio, kun UUID til BLE-beaconet de bærer. Denne UUIDen blir kun knyttet til navn i huben, og bare den som har tilgang til systemet kan koble UUID og person sammen.

7.3 Sikkerhetstrusler aktuelle for prototypen

Ethvert elektronisk system som angir posisjon og status på personell er utsatt for både passive og sikkerhetstrusler. En relevant passiv sikkerhetstrussel er at et trådløst system som sender data over eteren kan bli oppdaget av en fiendtlig aktør i det elektromagnetiske spekteret [7]. Dette kan føre til at posisjonen til fartøyet blir avslørt på bakgrunn av strålingen som emitteres. Ved å benytte Bluetooth og Zigbee-protokollen i prototypen kan man argumentere for at denne sikkerhetstrusselen ikke er relevant. Dette er i hovedsak på grunn av at rekkevidden på disse protokollene er veldig kort, og man vil mest sannsynlig bli oppdaget visuelt lenge før man blir oppdaget i det elektromagnetiske spekteret. Det er også verdt å nevne at sikkerheten til personell har førsteprioritet i fredstid, noe som gjør at et system som eventuelt har en høyere signatur i det elektromagnetiske spekteret, kan nyttes i fredstid.

En annen relevant passiv sikkerhetstrussel er at en fiendtlig aktør kan hente ut informasjon fra et slikt system, og dermed svekkes konfidensialiteten til systemet [10]. Likevel er det flere aspekter som gjør at dette er krevende ved den utviklede prototypen. For det første må man være innenfor kort avstand av fartøyet for å kunne av lytte signalene. Skulle man mot formodning være innenfor avstand til å avlytte systemet er det flere andre sikkerhetstiltak man må gjennom. Begge protokollene støtter AES-128 kryptering, noe som gjør det noe mer krevende å hente ut informasjon, men ikke umulig. I tillegg til å bryte krypteringen må man ha en programvare som kan nyttiggjøre seg av de ulike RSSI verdiene og UUIDen. Av denne grunn kan det argumenteres for at sannsynligheten for at konfidensialiteten til prototypen blir brutt er lav.

Et annet aspekt som er sentralt å diskutere er hvilke nytteverdier en eventuell fiendtlig aktør vil ha av informasjon knyttet til prototypen. Dersom en fiendtlig aktør kommer så nært innpå fartøyet vil det være andre ting som er av større interesse. Videre er det mulig å slå av eller la være å benytte systemet til kai der dette er en reell trussel. Det mest kritiske vil i hovedsak kunne være om en fiendtlig aktør finner tilstanden til fartøyet og personopplysninger. Når det kommer til personopplysninger, vil man knytte UUID til navn eller stilling i huben. Dette vil ikke sendes over nettverket og av den grunn kan man si at det vil være enda mer krevende å hente ut denne informasjonen. Når det kommer til muligheten for å kunne hente ut informasjon fra andre systemer kan dette unngås ved å sørge for at personellkontrollsystemet er fullstendig frakoblet og fysisk separert fra andre systemer om bord.

En aktiv sikkerhetstrussel et trådløst system som sender data over trådløse medier kan bli utsatt for er jamming, eller andre metoder en fiendtlig aktør kan benytte seg av for å sette systemet ut av spill [17, 32]. Tilgjengeligheten til systemet vil bli svekket hvis systemet slutter å fungere, enten grunnet fiendtlig påvirkning eller av andre grunner [10]. Kommunikasjonen i prototypen mellom BLE-beacons og HM-10 modulene vil bli lite påvirket av jamming. Årsaken til dette er at Bluetooth protokollen benytter seg av frekvenshopping spredt spektrum. I tillegg til dette benytter Bluetooth Adaptiv Frekvens Hopping, en teknikk som bytter frekvens hvis det er mye støy på den. Dette er i hovedsak for å unngå å benytte frekvenser som har mye trafikk, men gir også redundans mot jamming. Selv om Zigbee-protokollen på XBee-modulene har ikke samme robusthet mot jamming, er ikke jamming den største trusselen til prototypen. Elektronisk personellkontrollsystem blir nemlig beskyttet av korvetten selv. Fartøyet i praksis blir en Faraday-bur, som beskytter kommunikasjonssystemer inn i fartøyet fra utvendig påvirkning.

En annen aktiv sikkerhetstrussel er at et trådløst system som sender data over trådløse medier kan bli manipulert av en fiendtlig aktør og dermed påvirke integritet av systemet. Selv om dette er en relevant sikkerhetstrussel, vil den trolig være lite relevant for prototypen. Dette er på grunn av at det vil være vanskelig for en fiendtlig aktør å infiltrere systemet da de må være svært nære for å være innenfor dekningen til systemet.

8 Konklusjon

Helge Ingstad-ulykken i 2018 viste at Marinen opplever problemer med å oppnå personellkontroll om bord på sine fartøysklasser på en effektiv måte. I denne artikkelen viser vi en prototype for en elektronisk personellkontroll systemet, som tilfredsstiller tekniske krav i Skjold-klassen, og Marinens generelle bruker krav. I tillegg, er systemet laget for å tilfredsstille juridiske kravene for systemer som handler personopplysninger. Selv om artikkelen fokuserer Skjold -klassen, vil funnene ha føringsverdi til andre fartøystyper i Marinen.

Referanser

1. Amazon: Blue Charm Beacons, https://www.amazon.com/Blue-Charm-Beacons-BluetoothBC011MultiBecon/dp/B085XN9B7N/ref=sr_1_3?dchild=1&keywords= Bluetooth+Beacon&qid=161953875&sr=8-3, aksessert 10. apr. 2021.
2. Amazon: DSD TECH HM-10 Bluetooth, https://www.amazon.com/DSD-TECH-Bluetooth-iBeacon-Arduino/dp/B06WGZB2N4, aksessert 10. apr 2021.
3. Apple inc.: iBeacon, https://developer.apple.com/ibeacon/, aksessert 29. apr 2021.
4. Arduino: Language reference, https://www.arduino.cc/reference/en/, aksessert 1. Apr. 2021.
5. Arduino: UNO, https://www.arduino.cc/en/Main/arduinoBoardUno, aksessert 30. apr. 2021.
6. Arstad, S: Forsvarets forum. https://forsvaretsforum.no/nyhetsvarsel-sjoforsvaret/kristian-9-laget-knm-glimt-av-pepperkake/175101, aksessert 25. mai 2021.
7. CRFS: Naval Emission Control, https://www.crfs.com/applicationstory/naval-emcon-emissions-control/, aksessert 29. jan. 2021.
8. Datatilsynet: Behandlingsgrunnlag, https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/behandlingsgrunnlag/veileder-om-behandlingsgrunnlag/, aksessert 17. mar. 2021.
9. Datatilsynet: Grunnleggende personvernsprinsipper. https://www.datatilsynet.no/rettigheter-og-plikter/personvernprinsippene/grunnleggende-personvernprinsipper/, aksessert 17. mar. 2021.

10. Datatilsynet: Informasjonssikerhet og internkontroll, https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/informasjonssikkerhet-internkontroll/etablere-internkontroll/iverksette-styringssystem-for-informasjonssikkerhet/, aksessert 16. mar. 2021.
11. DIGI: Documentation XBee-PRO S2C DigiMesh® 2.4, https://www.digi.com/ resources/documentation/digidocs/pdfs/90001506.pdf, aksessert 15. apr. 2021.
12. DIGI: xbee, https://www.digi.com/xbee, aksessert 15. apr. 2021.
13. DIGI-Key: XBee S2C, https://www.digikey.com/en/product-highlight/d/digi-intl/xbee-s2c-802-15-4-rf-modules, aksessert 10 apr. 2021.
14. ELFA DISTRELEC: Mikrokontrollerkort, Uno, Arduino. https://www.elfadistrelec.no/ aksessert, 31. apr. 2021.
15. Fismen, T., Reiming, A.: Personlig samband/tracker-løsning på fartøy. Bacheloroppgave, Forsvarets Høgskole, Cyberingeniørskolen, Lillehammer, Norge (2021).
16. Frenzel, L.E.: Principles of electronic communication systems. 4th edn. McGraw-Hill Education, (2016).
17. Jaatun, M.G.: Sikkerhet uten en tråd, https://infosec.sintef.no/informasjonssikkerhet/2018/10/sikkerhet-uten-en-trad/, aksessert 15. apr. 2021.
18. Jinan Huamao Technology: Bluetooth, http://www.jnhuamao.cn/bluetooth.asp, aksessert 1. jan. 2021.
19. Jin Huamao Technology: Bluetooth 4.0 BLE module. http://www.jnhuamao.cn/bluetooth40_en.zip, aksessert 1. mar. 2021.
20. LOVDATA: Lov om nasjonal sikkerhet. https://lovdata.no/dokument/NL/lov/2018-06-01-24, aksessert 17. mar. 2021.
21. LOVDATA: Personopplysningsloven. https://lovdata.no/dokument/NL/lov/2018-06-15-38, aksessert 17. mar. 2021.
22. Malajner, M., Planinsic, P., Cucej, Z., Benkic, K.: Using RSSI value for distance estimation in wireless sensor networks based on ZigBee. In: 15th International Conference on Systems, Signals and Image Processing, pp. 303-306, IEEE, Bratislava, Slovak Republic (2008).
23. Matre, J.: Skipssjefen på KNM «Helge Ingstad»: Slik opplevde han det dramatiske havariet. https://www.vg.no/nyheter/innenriks/i/vmBGO4/skipssjefen-paa-knm-helge-ingstad-slik-opplevde-han-det-dramatiske-havariet, aksessert 8. feb. 2021.
24. Rapp,A.: xbee-arduino, https://github.com/andrewrapp/xbee-arduino, aksessert 7. apr. 2021.
25. Raspberry Pi Foundation: Raspberry Pi 4 model b, https://www.raspberrypi.org/products/raspberry-pi-4-model-b/, aksessert 17. mar. 2021.
26. ReelyActive: Bluetooth Low Energy (BLE) Identifier Reference, https://reelyactive.github.io/ble-identifier-reference.html, aksessert 25. mar. 2021.
27. Reinsnes, L.A., Anders Imenes A., Utvikling av støtteverktøy for nettverksbasert ESM. Bacheloroppgave, Forsvarets Høgskole, Cyberingeniørskolen, Lillehammer, Norge (2019).
28. RS: Raspberry Pi Display Kit, https://no.rsonline.com/web, aksessert 10. apr. 2021.
29. Sørensen, A., Egeland, E.S.: Agile Systemutviklingsmetoder. Masteroppgave, Universitetet i Agder, Kristiansand, Norge (2007).
30. Virgillito, D.: Tracking technologies and ther impact on privacy, https://resources.infosecinstitute.com/certification/3-tracking-technologies-and-their-impact-on-privacy/, aksessert 21. jan. 2021.
31. Vo, T.: Indoor position tracking based on arduino, XBee and ethernet shield, https:// github.com/thovo/Arduino-Indoor-Position-Tracking/blob/master/reports/Report.pdf, aksessert 21. jan. 2021.
32. Wilkins.S.: Wireless lan security threats. https://www.pluralsight.com/blog/it-ops/wireless-lan-security-threats, aksessert 22. mar. 2021.

Formalizing swarm security

Martin Strand

FFI, martin.strand@ffi.no

Abstract. Cheap, capable processors are becoming readily available, and these are being deployed to a wide variety of applications, ranging from internet-enabled refrigerators to swarms of specialised drones. As often is the case with innovations, security lags behind. We aim to introduce the fields of cryptography and swarm modelling to each other, and present a selection of security definitions and existing schemes that satisfy these. Such schemes are seen in context of different scenarios to increase security where it matters. To this end, we further develop a swarm model to identify when different key structures should be used.

Keywords: adversarial models · autonomy · lightweight

1 Introduction

For small, autonomous devices, the path from novelty to household item has been short. Anyone can come up with a potential application for a small sensor, smart device, drone or even underwater vehicles. Some of these applications are so vital or sensitive that the security requirements should be correspondingly strong. However, previous surveys [9,19] have made it clear that the foundation for secure designs for wireless sensor networks (WSN), internet of things (IoT) or mobile ad-hoc networks (MANET) is in its infancy.

The challenge in this domain is the multitude of constraints that may apply to our devices: battery life, computing performance, weak transmitters, low bandwidth, short transmission windows, limited storage, and so on. Many of the problems are solved for ordinary networks, but those solutions may not be applicable for this setting. For example, asymmetric cryptography is often too computationally expensive.

The intention of this work is to contribute to bridging the gap between real-life constraints and cryptography theory by building on rigid definitions from the latter, and suggesting ways these definitions could be satisfied while acknowledging the particular and unique limitations that make it hard to deploy existing security practices to small devices. We want to introduce two exciting fields to each other. In one direction by curating and presenting highly successful cryptographic models and definitions to a wider security audience. In particular, we have tried to *not* provide unecessary novelty in this respect. For the other direction, by keeping a critical eye to those definitions as they meet the world of practical considerations, this work is intended to be a part of the feedback loop to the theoretical side. The presentation is also intended to reach both cryptographers and practitioners in the domain of autonomous devices. We have kept the formalism of the original definitions even though we are not using them to prove concrete theorems in this work. The purpose of that is to provide a complete reference when designing constrained systems.

Our primary contribution is an à la carte menu of security notions an autonomous system could aspire to. We do not wish to add to the number of competing definitions. Instead, we have chosen well-established definitions from cryptography to fit reasonable requirements one could ask from such systems. We also discuss compatibility between the notions.

A secondary contribution is a refinement to how one can model a swarm of autonomous devices using graph theory. This is again a useful tool to characterise how the different definitions should be applied. In addition, we provide examples of how our contributions fit with existing solutions and scenarios.

The third pillar of this work is an introduction to challenges facing systems which may be deployed with secrets into an adversarial environment, but are denied any form of communication. This implies a theoretically impossible problem: encryption is worthless when the key is stored next to the ciphertext. However, we show that there is some leeway once one considers time as well.

1.1 Adversarial capabilities

One cannot assess security without making assumptions about the adversary. We generally assume the existence of a computationally bounded, active and adaptive adversary. For non-cryptographers, that means:

- The adversary can spend at most a polylogarithmic amount of time and space on computations, with respect to size of the keyspace. Conversely, an unbounded adversary would for instance be able to try all keys from a key space, and choose the correct one.
- An active adversary controls the network, and may read, inject, drop and modify messages on the network. In contrast, a passive adversary may only read the traffic.
- The adaptivity means that the adversary can corrupt parties at will throughout the execution, and then control their actions completely. A static adversary has to choose the set of corrupted parties before the protocol starts. All parties (or instances of such) have a freshness flag which is set to false when corrupted. We require that the final attack can only be mounted against parties that are still fresh.

On the other hand, we do not make any assumptions on what the adversary is trying to achieve. Loosely speaking, the adversary wins if it can make the system behave in any other way than defined for each notion. The observant reader will see that we have given the adversary the capability to drop all messages, and that automatically makes the adversary able to run effective denial-of-service (DoS) attacks. In order to avoid pathological cases like that, cryptographers have designed a paradigm where we challenge the adversary to a game for each security goal, and the adversary will win if it does better than a random algorithm.We will later list the specific goals for the adversary to attack.

1.2 Related work and IoT security

Previous work [9, 19] has surveyed the existing state of security modelling for WSNs, IoT, and MANETs, with discouraging results. Do, Martini and Choo note that "[o]ther security-based research should look to cryptographic protocols as the gold standard for adversary models (...)", and add: "IoT security, particularly, is a research field in its infancy." The earlier surveys cite a number of papers that in total give the impression that even routine use of authentication is lacking in many applications.

According to Silvio Micali in his IACR Distinguished Lecture at Crypto 2020, models might be one of cryptography's strongest assets. Starting from mission goals, and assumptions about the environment and adversarial capabilities, one can formulate security requirements and reason about these. From the literature we have previously reviewed, only two works are worth mentioning here. Sen [18] gives a comprehensive overview of the field, and lists a number of security and functionality requirements. The requirements are well-arranged: they formulate guarantees one wants to make, rather than describing the inner workings of a given system. Akram et al. [2] introduce an intriguing system with corresponding security requrements. However, the requirements are in much greater extent tied to the authors' suggested system.

1.3 Organisation of the paper

The paper is structured as follows. The following section introduces a new variant of how to model a swarm. The model takes into account that the topology can change over time. Section 3 introduces relevant security goals, of which application to the model is further discussed in Section 4. We then follow up in Section 5 with three different examples of how one can prioritise the definitions in practical cases. In Section 6, we discuss the related problem of protecting data at rest in cases with a high risk of adversarial corruption. Finally, we conclude in Section 7 and give an overview over some of the numerous open problems regarding strong security under these challenging conditions.

2 What is a swarm?

The mission of cryptography can be formulated as ensuring that functionalities can work even in the presence of a powerful adversary. In light of this view, a secure channel has the same functionality as any reliable channel, but with the extra property that Eve or Mallory are unable to eavesdrop or manipulate the data without detection. Similarly, one can phrase our task here as taking the definition of a swarm, and ensuring that the definition could be fulfilled, even in an adversarial environment.

However, when researching a suitable definition of a swarm, we came across the following statement by Hamann: "It is interesting to notice that there seem to be no explicit definitions of swarms in the literature" [12, Sec. 1.1.1]. We cannot state precisely what a swarm is, but we can at least model it, with some inspiration from Hamann. In his book, he describes a random graph $G = (V, E)$ as a viable model for swarms, where the nodes represents the devices and each edge represent a communication connection. We refine the model slightly, by instead applying directed random geometric graphs, first described by Michel et al. [14]. The advantage is that we can model devices with non-uniform sending and receiving levels. We present a simplified version, which generalises random geometric graphs in the natural way, but for which some of the results by Michel et al. may not hold. In particular, we do not explicitly assume that the different radii are distributed according to a Pareto distribution.

Definition 1. *A* directed random geometric graph *is a graph $G = (V, E)$ such that each $v \in V$ is a point in $\mathbb{R}^n$ and satisfies the following:*

1. *The nodes are randomly sampled from some distribution on a (potentially bounded) region of $\mathbb{R}^n$.*
2. *For each node $v \in V$ there exists a radius r_v.*
3. *Let $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a metric. There exists an edge $e \in E$ for each pair of nodes (v, w) with $d(v, w) < r_v$.*

The graph will model the communication capabilities in the collection of devices at a given time t. Let $v \in V$ be a node and define the following two sets relative to v:

- $N_v^t = \{u \in V \mid (v, u) \in E\}$ is the neighbourhood around v at time t, which contains the nodes that can receive any messages directly from v.
- Define u to be reachable from v if there exists a path from v to u, and let $\bar{N}_v^t = \{u \in V \mid u \text{ reachable from } v\}$, which we will call the swarm around v at time t.

One can view the graph as a representation of a relation, and then let $\bar{N}_v^t$ be the transitive closure of N_v^t. This last item states that if two devices cannot even communicate through proxies, then they are effectively not in the same swarm at that moment.

Remark 1. We believe that this asymmetric definition may be a realistic model of real challenges regarding devices in challenging environments. For instance, some high-powered devices may be able to reach the whole network, but only receive from its closest neighbours. However, we hypothesize that for most real-life situations, the graphs will be undirected. This will simplify many aspects that follows, but the reader should keep in mind that one should consider the corner cases properly when designing protocols.

Next we must discuss the size of a swarm. Hamann [12] suggests between 10^2 and 10^{23}, beyond which one should use statistics to analyse the system as one would for a gas. However, Hamann remarks, one can also consider swarms of three members, and for the purpose of our work: two will also be applicable.

Looking ahead, our model will allow us to create three different scenarios, depending on the size of the swarm and the computational resources of the devices. The scenarios provide a trade-off between complexity, security and resources.

3 Security goals

Next, we provide a series of potential security goals. We stress *potential*, because this list is intended to be used *à la carte*, depending on the concrete application.

When instantiating any of these notions, one must choose a target security level and use that as input to the key generation algorithm. This security parameter is of particular importance to record the potential loss of security that may come from reductions in proofs. This paper only contains one such proof, and as that proof only contains a standard argument that halves the advantage, we have opted to omit explicitly mentioning the security parameter to increase readability.

3.1 Channel privacy

Our first definition is just the normal definition for IND-CCA2, message indistinguishability under adaptive chosen ciphertext attacks.

Definition 2 (Ciphertext indistinguishability, IND-CCA2). The adversary should not be able to read the contents of a message in the channel. *Let* $\mathcal{ES} = (\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ *be an encryption scheme.*

1. *The challenger runs* $\mathcal{ES}.\mathsf{KeyGen}$ *and sets up two oracles* $\mathcal{O}^{\mathsf{Enc}}, \mathcal{O}^{\mathsf{Dec}}$. *It then chooses a bit* b *at random, and sends a signal to the adversary* $\mathcal{A}$.
2. $\mathcal{A}$ *may send a large number of queries to* $\mathcal{O}^{\mathsf{Enc}}, \mathcal{O}^{\mathsf{Dec}}$. *Eventually, the adversary submits two messages* m_0, m_1 *of identical length.*
3. *The challenger encrypts* $c = m_b$, *and returns it to the adversary.*
4. *The adversary may again query the oracles, except for decryption oracle calls on the ciphertext* c. *Finally, the adversary submits a bit* b', *and wins if* $b = b'$.

The scheme $\mathcal{ES}$ *is IND-CCA2-secure if the adversary's advantage*

$$\mathsf{Adv}^{\textit{ind-cca2}}_{\mathcal{ES}}(\mathcal{A}) = \left| \Pr[b = b'] - \frac{1}{2} \right|$$

is negligibly small.

The definition of a negligible value or function may depend on the application. For asymptotic arguments it is defined as a function eventually being smaller than $1/p(\lambda)$ for any polynomial p in the security parameter λ. For example $1/2^{\lambda}$, is a negligible function. On the other hand, concrete systems may define anything smaller than a given constant as negligible.

3.2 Swarm privacy

Now we discuss what we mean by swarm privacy. A swarm may consist of several devices that leave and join during the mission. This problem is related to that of broadcast encryption and multicast encryption [10].

The fundamental goal is that any message sent within a swarm should be readable for any active member, but not for outsiders and lost devices. This means we have to be able to handle join and leave operations while the system is online. An important question is how one can decide which devices should be considered lost, and henceforth apply the leave operations on those. A more complicated question is that of *who*: is this an individual decision, a swarm consensus, or a question of hierarchy and hence trust?

Remark 2. The reader may at this point compare this problem to that of group chats, and consider group key exchange protocols. Such protocols are outside the scope of this work, as they demand too much network traffic, constant liveliness, and expensive computations. However, it is worth mentioning a recent preprint by Weidner et al. [20], which aims to provide decentralised key agreement for large groups. A special emphasis is placed on post-compromise security (PCS) and forward secrecy (FS). Weidner et al. are discussing the difficult problem of how one should reach consensus about membership.

Definition 3 (Swarm privacy). Only active, honest players may read broadcast messages. *We set up a game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.*

- *The challenger runs the setup and key generation algoritms for a system of n users. It also chooses a bit b uniformly at random.*
- *The adversary may query the oracle multiple times, each time one of the following queries*
  $\mathsf{corrupt}(i)$ *The key material held by user i is revealed to $\mathcal{A}$.*
  $\mathsf{Enc}(S, m)$ *The message m is encrypted with the members of a subset S of $\{1, ..., n\}$ as intended receipients.*
  $\mathsf{Dec}(S, c)$ *If S is the correct audience for the ciphertext c, then the oracle outputs the decryption m.*
- *The adversary selects two messages m_0, m_1 of equal length and submits them to $\mathcal{C}$ along with a set S^*. The challenger verifies that S^* does not contain any user previously corrupted and only then responds with $c' = \mathsf{Enc}(S^*, m_b)$.*
- *The adversary may again provide new queries with two restrictions:*
  - *any query $\mathsf{corrupt}(i)$ is on the condition that $i \notin S^*$, and*
  - *the adversary may not query $\mathsf{Dec}(S^*, c^*)$.*
- *The adversary submits a bit b' and wins if $b = b'$.*

We define the advantage of $\mathcal{A}$ as

$$\mathsf{Adv}^{swarm}(\mathcal{A}) = \left| \Pr[b = b'] - \frac{1}{2} \right|$$

and say that the system is secure if $\mathsf{Adv}^{swarm}(\mathcal{A})$ is negligible.

This definition is intended to capture the essence of two previous definitions by Gentry and Waters [11], and Panjwani [16], which were specific to the asymmetric and symmetric cases respectively. We return to a more thorough discussion of similarities and differences later.

3.3 Authenticity

A swarm may be used to observe a large area and provide its sightings to a central intelligence service. Recall that we assume that the adversary is able to inject messages into the network. The adversary can then try sending false reports to lure the others. We therefore want swarm members to be able to attribute messages to their senders. This is known as authenticity, and is captured by recalling the definition of ciphertext integrity (INT-CTXT):

Definition 4 (Ciphertext integrity, INT-CTXT). An encrypted message should be received as intended by its sender. *Let* $\mathcal{ES} = (\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ *be a symmetric encryption scheme and let* $\perp$ *denote error. We set up the following experiment* $\mathsf{Exp}^{\textit{int-ctxt}}_{\mathcal{ES}}(\mathcal{A})$*:*

- *Generate a key* $K \leftarrow \mathsf{KeyGen}$*, and set* $S = \varnothing$*.*
- *Give the adversary access to encryption and decryption oracles* $\mathcal{O}^{\mathsf{Enc}}, \mathcal{O}^{\mathsf{Dec}}$*, with the additional programming that:*
  - *whenever the oracle* $\mathcal{O}^{\mathsf{Enc}}$ *responds with some* c_i *when queried on some message* m_i*,* S *is updated as* $S = S \cup \{c_i\}$
  - *when* $\mathcal{A}$ *queries* $\mathcal{O}^{\mathsf{Dec}}$ *with* c*, let* m *be the output from decrypting* c*. If* $m \neq \perp$ *and* $c \notin S$*, halt and output 1*

  The encryption scheme $\mathcal{ES}$ *is INT-CTXT-secure if*

$$\mathsf{Adv}^{\textit{int-ctxt}}(\mathcal{A}) = \Pr[\mathsf{Exp}^{\textit{int-ctxt}}_{\mathcal{ES}}(\mathcal{A}) \textit{ outputs } 1]$$

is negligible for all adversaries $\mathcal{A}$*.*

Note, however, that this definition does not protect against replay or reordering attacks. For an example of a stronger notion, see Bellare, Kohno and Namprempre's stateful definition INT-sfCTXT [4]. Furthermore, it does not tie the concept of "sender" to a device identification. Such binding may be done implicitly by a key.

3.4 Anonymity

We also want messages to remain anonymous so that any adversary monitoring the network can do no better than traffic analysis to assess who sent which message. We have reviewed two works that investigate this problem for symmetric encryption, by Chan and Rogaway [6] (anonymous nonce-based authenticated encryption, anAE), and Banfi and Maurer [3] (probabilistic authenticated encryption, pAE). The latter considers the special case of probabilistic encryption, while the first is more general, but has to add more machinery around the encryption scheme in order to satisfy their own definition. Nonetheless, we opt for the more general definition, as it is closer to the common use of symmetric encryption. We give an informal presentation of the notion here, and refer to the original work for the details[1].

[1] We justify this omission for two reasons: The whole description with proper context fills a complete section, and our presentation would not improve on Rogaway and Chan's Figure 1, which we highly recommend to the reader.

The adversary is challenged to distinguish between two games: one implementing the real cryptosystem Π, and one implementing an ideal functionality.

The key difference between the two is that the ideal functionality replaces the encryption oracle with one that returns a random string of equal length, while storing the original request. This ensures that the ideal functionality provides both confidentiality and privacy, since the ciphertext is independent of both the plaintexts and the identity. The ideal decryption answers with the error symbol $\perp$ unless the decryption request perfectly matches a record stored by the encryption service, and the nonce is within some accepting policy Nx. The adversary wins if it outputs 1 while interacting with the real game, and loses otherwise.

Security is then defined as the quantity

$$\mathsf{Adv}^{\text{anae}}_{\Pi,\mathsf{Nx}}(\mathcal{A}) = \left|\Pr[\mathcal{A}^{\text{real}^{\text{anae}}_{\Pi,\mathsf{Nx}}} \rightarrow 1] - \Pr[\mathcal{A}^{\text{ideal}^{\text{anae}}_{\Pi,\mathsf{Nx}}}] \rightarrow 1\right|.$$

Interestingly, Rogaway and Chan's concept of nonce policies can also be used to avoid replay attacks.

3.5 Topology hiding

As well as keeping each player anonymous, we would like to hide the network topology from the adversary. We adapt a definition by Moran, Orlov and Richelson [15] to our concrete case. The definition is based on the assumption that any node can detect all neighbours within its range, i.e. count its edges. Notice that the following formulation is static in the sense that the adversary has to choose which players to corrupt early in the game. This seems unavoidable, as adaptive corruptions would essentially give the adversary a way to traverse the graph.

Recall that $G = (V, E)$ is a directed graph, and that N_v^t is the neighbourhood around v at time t. Let $G' = (V, E')$ be the associated undirected graph such that for all $v_1, v_2 \in V$, let $(v_1, v_2) \in E'$ if either $(v_1, v_2) \in E$ or $(v_2, v_1) \in E$. We fix (and therefore omit) the time t for now, due to the non-adaptive nature of this definition.

Definition 5. *Let $\mathcal{G}$ be a set of undirected graphs with at most n nodes, and let Π be a protocol capable of running on all $G \in \mathcal{G}$. Each player $P_1, \ldots, P_N$ is equipped with key material $k_1, \ldots k_N$.*

- *$\mathcal{A}$ chooses a corrupt subset S and learns the key material for all players $P \in S$ and two graphs $G_0 = (V_0, E_0), G_1 = (V_1, E_1)$ such that $S \subseteq V_0 \cap V_1$ and $N_{P,G_0} = N_{P,G_1}$ for all $P \in S$, i.e. that all of the corrupted nodes have equal neighbourhoods in both graphs. $\mathcal{A}$ sends (S, G_0, G_1) to the challenger.*
- *The challenger chooses a random bit b and runs Π on G_b. It interacts with $\mathcal{A}$ for all traffic involving $P \in S$.*
- *Finally, $\mathcal{A}$ outputs a bit b', and wins if $b = b'$.*

The protocol Π is indistinguishable under chosen topology attack *(IND-CTA secure) over $\mathcal{G}$ if the quantity*

$$\mathsf{Adv}_{\pi}^{ind\text{-}cta}(\mathcal{A}) = \left|\Pr[b = b'] - \frac{1}{2}\right|$$

is negligible.

The adversary sees all its neighbours at the physical layer. The consequence of fixing the time is that one may be unable to guarantee anonymity if the network topology changes. Honest parties beyond the range of the adversary may change without impacting the gist of the definition. We have previously hypothesized that most graphs will be undirected in practice. This definition will apply in those cases. Applications must consider these limitations.

4 Applying the definitions to the swarm models

We will now bind the security requirements to our model of a swarm. Recall that we defined the two sets N_v^t and $\bar{N}_v^t$, which are the devices that are directly within range of v, and those reachable through relays. We defined the latter as the swarm around v. Due to our restriction to constrained devices, we only consider symmetric keys, and consider a device a member of the swarm if it has a valid key.

From this, we consider three scenarios:

Individual keys All devices share keys pairwise. Hence, every message must be sent to all other devices individually. This might be beneficial for very small swarms with low communication rate and high bandwidth, but limited resources to negotiate group keys. In this case, N_v^t and $\bar{N}_v^t$ exist only for bookkeeping.

Local group keys The node v maintains a group key for the members of N_v^t, and updates it if $N_v^t \neq N_v^{t+1}$. Limited group keys might reduce the consequence if a device is lost, and may be a favourable choice for larger networks with low mobility and a strong adversarial threat.

Global group keys The node v maintains a group key for $\bar{N}_v^t$ and updates if necessary. If $\bar{N}_v^t = \bar{N}_w^t$ for all $w \in \bar{N}_v^t$, then a fixed v may if necessary act as a key centre for the swarm.

For each of these, we must handle two operations: join and leave. This single sentence is worth a series of papers by itself. A device may join if it can produce valid credentials, i.e. having been keyed appropriately. Depending on the above scenario, appropriate rekeying of the other devices in the swarm may follow. Leave, on the other hand, is only somewhat easy if the device itself announces its intention of parting. If the device leaving was still trusted, it could just be asked to delete its keys, and the swarm would not have to replace keys held by its former member. However, the leave operation should imply that the device is no longer trusted, and quite possibly already have been corrupted by the adversary.

The conditions on when the swarm should initiate a forced leave is outside the scope of this work. We also acknowledge the multitude of problems connected to just the four words "when the swarm should", but have to postpone those to future work.

Furthermore, we need to demonstrate whether it is feasible to satisfy the security requirements we have stated above for the different scenarios.

4.1 Individual keys

We start with the easiest case. Since all neighbours share individual keys, there is no group, and so Definition 3 and Definition 5 are vacuously satisfied. For individual keys, join and leave is handled implicitly. One can choose to do key exchange during operation, or pre-key the devices for all admissible pairs.

To satisfy channel privacy and device accountability we can use any standard authenticated encryption, and the anonymity requirement can be satisfied by using Chan and Rogaway's protocol [6].

4.2 Group keys

We treat local and global group keys together. They differ in scope, but not necessarily in technique. To handle join and leave, we suggest using the protocol of either Gentry and Waters [11], or Panjwani [16]. The former describe a public key protocol: to send a message using the protocol, one specifies a set S of recipients and encrypts it using a function of every recipent's public key. A single private key is sufficient (and necessary) to decrypt. We refer to the original paper for the technical details. This protocol allows the group to distribute a new key which can again be used for communication until the next topology change. The authors prove that their protocol satisfies a security definition similar to ours.

Proposition 1. *The Gentry-Waters system satisfies swarm privacy.*

The proof is a simple composition of two definitions and a conversion between real-or-random and left-or-right notions, and is included in the full version of the paper.

One apparent drawback with the work of Gentry and Waters is that it is based on a variant of the Diffie-Hellman problem and bilinear maps over elliptic curves. While this is considered secure today, it will not be able to stand against a quantum computer, and so there is a need to reinstantiate their concepts using post-quantum techniques.

For the symmetric case, one can use Panjwani's improvement to the Logical Key Hierarchy protocol [13]. The idea is that one builds a binary tree of keys where each leaf node represents a device, and the root node holds the group key for the complete set. All devices know the keys on the path between themselves and the root node. In order to send a message (say, a group key) to a subset S of the leaf nodes, one must choose the minimal set of keys held by members of

S, but such that for any node v not in S, no key in the path between v and the root is included.

The tree must be maintained by a key centre, which is then responsible for distributing new node keys for each topology change, and a group key to the active devices. Definition 3 is an adaptation of the definition used by Panjwani.

In sum, these two approaches provide join/leave functionality. However, what they have in common is that the topology is partly leaked every time it is invoked, and any system employing either of these techniques cannot hope to be topology hiding over time. Further research is needed to see if one can reconcile topology hiding and group encryption.

5 Example applications

We can now apply our work to some real-world examples, for each highlighting their unique limitations. Examples like these can be composed across domains with the help of relay nodes.

5.1 Submarine communications

Consider a small set of autonomous underwater vehicles (AUV). Weight may not be the primary issue, so we can assume sufficient computing abilities. However, every transmission is through an acoustic channel whose bandwidth may only be around 500 bit/s. This requires the overhead to be as small as possible. This exact problem was analysed and tested by Dini and Duca [8]. Their work fits our model very well:

- In the simple case, there is a gateway g with $N_g^t = V$, while for all other nodes v, $N_v^t = \{g\}$. Hence, nodes communicate directly with the gateway, which in turn forwards the message to the intended receiver.
- Channel privacy is provided by AES used in CBC-CTS mode to avoid overhead.
- The authors use a standard message authentication code (MAC) to provide integrity. However, due to the low bandwidth, they truncate it down to 32 bits, noting that the same bandwidth makes an online attack extremely time-consuming. Hence, the adversary's success probability is limited not in having an overwhelmingly large denominator, but by realising that the numerator is bounded by the physical surroundings. This can be modelled by restricting the number of decryption oracle queries in the security game.
- Swarm privacy can be satisfied due to their group key distribution and revokation. Their key distribution is closely related to that of Panjwani.
- Topology hiding is not an issue considered by Dini and Duca. Recall also that the methods we discussed earlier are incompatible with this notion.

5.2 A swarm of drones

Small, airborne drones have a completely different set of limitations. The bandwidth is high, but the onboard computer should be small and use as little energy as possible on computations.

A key feature of drone swarms is that its members can be replaced continuously, and we must assume that the devices shift their position relative to each other. We therefore find ourselves in a situation where group keys may be the right choice.

- Channel privacy and ciphertext integrity can be achieved by using a suitable authenticated encryption scheme.
- Swarm privacy follows the discussion of group keys, but is dependent on a robust mechanism for deciding which members should be excluded.
- Anonymity and topology hiding may or may not be important for this scenario, and may in fact be less important the more members of the swarm, as each drone becomes less crucial. Hence, we expect swarm privacy to take precedence.

In practice, such devices could be augmented with tamper-resistant cryptographic modules, in which one can place a reasonable level of trust.

6 Data at rest

Small, resource constrained devices do not only pose a challenge for communication. One should also expect that such devices could be captured in a working state and brought directly to a highly skilled laboratory: it may inject or extract any instruction or data through means that the original designer could use[2]. For this section, we no longer consider the graph of communicating devices, but focus on one particular, possibly isolated, device. We describe three informal security requirements, and then sketch how they could be satisfied.

Data authenticity Only data from authenticated functionalities should be accepted by the device.

Data obsoletion Any data no longer needed for the mission should remain unreadable.

Secure storage Data still needed for the mission should not be extractable by the adversary.

Data authenticity suggests that the data stored on the device should only originate from pre-loading, communication channels or sensors, all of which could sign the data before providing it to the on-board computer. The on-board computer should then only use the data in computations after verifying the authentication tags. While this may seem cumbersome for small devices, it has already been

[2] In essence, a statement from the manufacturer like “it is theoretically possible, but it would be very hard” here translates to “this lab can do it”.

tried on a drone with fascinating success [7]. The advantage would be – given certain assumptions – that the adversary could not insert malicious instructions or data on a device and then release it back into operation. Note that we do not consider the independent problem of the adversary exposing the sensors for misleading surroundings or physical attacks on the device.

Data obsoletion includes data that could be useful *after* the mission, e.g. sensor data stored on a device and only unloaded at a later stage. We propose two simple strategies: either using public key cryptography where only the public key is given to the device, or the perhaps more computationally friendly variant where the device is equipped with a symmetric key K_0. After each data object has been encrypted using some key K_i, replace K_i with $K_{i+1} = f(K_i)$ for some suitable one-way function f. Since K_i is deleted, all data protected by that key can now be considered cryptographically deleted until paired with another device that holds an K_j, $j \leq i$. This idea is closely related to that of forward secrecy.

Secure storage is the most difficult requirement of these. There is data that the device may need to access again and therefore must be able to decrypt. At a theoretical level, that means that the data could just as well have been unencrypted: a key next to a locked chest is effectively an unlocked chest. It becomes slightly more nuanced if one can include time in the picture.

In some sense, the main inspiration is again that of cryptographic deletion. If data only exists in an encrypted form, and the key is destructed, then the data can be considered deleted. In extension, data only comes into existence once the appropriate key is reconstructed.

Correspondingly, our definition does not consider data. Assume that the payload data has been logically structured by some directed graph, and that each part is encrypted with a new key. Identify each key with the corresponding node, so that reaching the node equals knowledge of the key, and hence the data.

Each node also contains information on all outgoing edges. Following the edge should be computationally costly, and preferably non-parallelisable. Concretely, every time the internal state reaches a node one can decrypt the corresponding data packet. It contains one puzzle for each edge, and the key for the other node of the edge can only be found by solving the puzzle. This idea motivates the following definition:

Definition 6. *Let $\mathcal{ES}$ be an encryption scheme. Let $G = (V, E)$ be a graph and $v_i \in V$. For each v_i define $L_i = (v_j, \mathsf{aux}_{v_j})$ as a list of all neighbouring nodes along with auxilliary data for each neighbour. A* Cryptographically Bounded Bandwidth Channel *consists of two algorithms* ($\mathsf{Setup}, \mathsf{RecoverKey}$) *with the following properties:*

Setup *An efficient algorithm that takes in a directed graph $G = (V, E)$ and a delay t, and outputs a related graph $(V', \varnothing)$. For each v_i, generate a key sk_i. Each node $v'_i \in V'$ corresponds to a node $v_i \in V$, and contains an encrypted list of neighbours of v_i, $\{\mathcal{ES}.\mathsf{Enc}(sk_i; L_i)\}$, and the corresponding auxilliary data.*

$\mathsf{RecoverKey}$ *Takes in the auxilliary data aux_{v_j}, and reconstructs the decryption key sk_j for v_j.*

The channel has a weakly (t, p)-bounded output rate if, with probability p, it takes at least tn time to recover n nodes. The channel has strongly (t, p)-bounded output rate if, with probability p, the time to recover each node is at least t.

Informally, view v' as an encrypted version of v and its edges. Implicitly, we assume that G is connected. If G has $n > 1$ components, one can simply consider the case of n instances of the channel.

We now present a simple, informal example to demonstrate how to use this definition. Assume Alice wants to navigate through an area without communicating with others, but wants to keep as much as possible of the map obscured at all times, yet get access every part of the map when necessary.

Divide the map into logical sections, and create a map graph: Each region is a node, and neigbouring regions are connected with edges. Assume that each area takes time t to traverse. Use the graph and time t as input to the setup phase, and use the keys generated in the definition to encrypt each region. Before starting the expedition, delete all keys but one, say sk_1, and set off in region 1.

Alice can see the labeling of the neighbouring regions, and choose one, say 2. In the time she needs to relocate to the next region, she reconstructs the corresponding key, and is able to decrypt the next map section once she reaches the border. She can then choose a new region to move into. Alice must now delete all data from region 1: the map, the key and the reconstruction data for other neighbours of region 1. She can only return to 1 if 1 is a neighbour of 2 (remember, G is a directed graph), or any of the other regions she will explore during her expedition.

This notion can be instantiated using time-lock puzzles [17], whose objective is to delay decryption. The problem is that such problems depend on CPU time (hence, a data center would have a large advantage relative to a resource-constrained device), whereas we would like a scheme that was able to keep the lock until a certain amount of wall-clock time had passed. This is an intrinsically hard problem given that we do not want to allow communication.

The best candidate so far is that of Abadi, Burrows, Manasse and Wobber [1]. In contrast to other candidates, Abadi et al. use memory latency to generate a problem. This number depends on physics and is relatively constant across platforms. The idea is to generate a problem such that the quickest way to solve it repeatedly is by generating a table of all solutions and do lookups for each instance. However, that table should be so large that it cannot fit in cache. For each instance, a new portion of the table must therefore be loaded from memory to cache, causing the delay.

The problem is typically to invert a hash function from a small domain. By repeatedly hashing and xor-ing the preimage with a counter value, one is forced to undo each step on the way. However, it has so far not proven suitable to generate a large key, and so we conclude that further work is needed, while also noting that continuous computations are less than ideal for a resource constrained device. In lack of a sound cryptographic solution, we note that tamper protection plays a crucial role in protecting data at rest.

7 Conclusion

We have presented a selection of security definitions suitable for swarms of devices with limited resources, possibly operating in hostile environments. Given that this is highly relevant for governments, military applications and also commercial interests, our hope is that this paper spurs a strengthened interest for the features modern cryptography has to offer the greater field of cyber security.

We have throughout this work pointed out loose ends and possibilities for further research. We conclude the paper by summarising these.

Forced leave The swarm needs a mechanism to determine which devices to be considered lost, and therefore exclude them. Given that we assume that the adversary might control some of our devices, how can we avoid the adversary trying to blame and exclude an honest (and possible crucial) device? In some settings one might consider special devices with elevated rights. Other times, one have to find a suitable approximation to a solution to the problem of Byzantine agreement.

Anonymity and swarm privacy Can there exist notions that combine efficient multicast with anonymity for any individual device in the network?

Quantum-safe broadcast and multicast Gentry and Waters' asymmetric multicast encryption scheme is based on classic cryptography. A quantum-safe instantiation is an open problem.

Keys and data on the same device Is it possible to delay decryption in such a way that it is at least highly correlated with offline wall-time?

References

1. Martín Abadi, Michael Burrows, Mark S. Manasse, and Ted Wobber. Moderately hard, memory-bound functions. *ACM Trans. Internet Techn.*, 5(2):299–327, 2005. doi:10.1145/1064340.1064341.
2. Raja Naeem Akram, Pierre-François Bonnefoi, Serge Chaumette, Konstantinos Markantonakis, and Damien Sauveron. Secure autonomous UAVs fleets by using new specific embedded secure elements. In *2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, China, August 23-26, 2016*, pages 606–614, 2016. doi:10.1109/TrustCom.2016.0116.
3. Fabio Banfi and Ueli Maurer. Anonymous symmetric-key communication. Cryptology ePrint Archive, Report 2020/073, 2020. https://eprint.iacr.org/2020/073.
4. Mihir Bellare, Tadayoshi Kohno, and Chanathip Namprempre. Authenticated encryption in SSH: Provably fixing the SSH binary packet protocol. In Vijayalakshmi Atluri, editor, *ACM CCS 2002*, pages 1–11, Washington, DC, USA, November 18–22, 2002. ACM Press. doi:10.1145/586110.586112.
5. Ran Canetti. Universally composable security: A new paradigm for cryptographic protocols. In *42nd FOCS*, pages 136–145, Las Vegas, NV, USA, October 14–17, 2001. IEEE Computer Society Press. doi:10.1109/SFCS.2001.959888.
6. John Chan and Phillip Rogaway. Anonymous AE. In Steven D. Galbraith and Shiho Moriai, editors, *ASIACRYPT 2019, Part II*, volume 11922 of *LNCS*, pages 183–208, Kobe, Japan, December 8–12, 2019. Springer, Heidelberg, Germany. doi:10.1007/978-3-030-34621-8_7.

7. Jung Hee Cheon, Kyoohyung Han, Seong-Min Hong, Hyoun Jin Kim, Junsoo Kim, Suseong Kim, Hosung Seo, Hyungbo Shim, and Yongsoo Song. Toward a secure drone system: Flying with real-time homomorphic authenticated encryption. *IEEE Access*, 6:24325–24339, 2018. doi:10.1109/ACCESS.2018.2819189.
8. Gianluca Dini and Angelica Lo Duca. A secure communication suite for underwater acoustic sensor networks. *Sensors*, 12(11):15133–15158, 2012. doi:10.3390/s121115133.
9. Quang Do, Ben Martini, and Kim-Kwang Raymond Choo. The role of the adversary model in applied security research. *Computers & Security*, 2018. URL: http://www.sciencedirect.com/science/article/pii/S0167404818306369, doi:https://doi.org/10.1016/j.cose.2018.12.002.
10. Amos Fiat and Moni Naor. Broadcast encryption. In Douglas R. Stinson, editor, *CRYPTO'93*, volume 773 of *LNCS*, pages 480–491, Santa Barbara, CA, USA, August 22–26, 1994. Springer, Heidelberg, Germany. doi:10.1007/3-540-48329-2_40.
11. Craig Gentry and Brent Waters. Adaptive security in broadcast encryption systems (with short ciphertexts). In Antoine Joux, editor, *EUROCRYPT 2009*, volume 5479 of *LNCS*, pages 171–188, Cologne, Germany, April 26–30, 2009. Springer, Heidelberg, Germany. doi:10.1007/978-3-642-01001-9_10.
12. Heiko Hamann. *Swarm Robotics - A Formal Approach*. Springer, 2018. doi:10.1007/978-3-319-74528-2.
13. Hugh Harney and Eric J. Harder. Logical Key Hierarchy Protocol. Internet-Draft draft-harney-sparta-lkhp-sec-00, Internet Engineering Task Force, April 1999. Work in Progress. URL: https://datatracker.ietf.org/doc/html/draft-harney-sparta-lkhp-sec-00.
14. Jesse Michel, Sushruth Reddy, Rikhav Shah, Sandeep Silwal, and Ramis Movassagh. Directed random geometric graphs. *Journal of Complex Networks*, 7(5):792–816, 04 2019. arXiv:https://academic.oup.com/comnet/article-pdf/7/5/792/30157011/cnz006.pdf, doi:10.1093/comnet/cnz006.
15. Tal Moran, Ilan Orlov, and Silas Richelson. Topology-hiding computation. In Yevgeniy Dodis and Jesper Buus Nielsen, editors, *TCC 2015, Part I*, volume 9014 of *LNCS*, pages 159–181, Warsaw, Poland, March 23–25, 2015. Springer, Heidelberg, Germany. doi:10.1007/978-3-662-46494-6_8.
16. Saurabh Panjwani. Tackling adaptive corruptions in multicast encryption protocols. In Salil P. Vadhan, editor, *TCC 2007*, volume 4392 of *LNCS*, pages 21–40, Amsterdam, The Netherlands, February 21–24, 2007. Springer, Heidelberg, Germany. doi:10.1007/978-3-540-70936-7_2.
17. Ronald L. Rivest, Adi Shamir, and David A. Wagner. Time-lock puzzles and timed-release crypto. Technical report, 1996.
18. Jaydip Sen. Security in wireless sensor networks. *CoRR*, abs/1301.5065, 2013. URL: http://arxiv.org/abs/1301.5065, arXiv:1301.5065.
19. Martin Strand and Jan Henrik Wiik. Kryptografisk sikring av autonome og ubemannede enheter – eksisterende forskning. FFI-rapport 19/02042, FFI, 2019.
20. Matthew Weidner, Martin Kleppmann, Daniel Hugenroth, and Alastair R. Beresford. Key agreement for decentralized secure group messaging with strong security guarantees. Cryptology ePrint Archive, Report 2020/1281, 2020. https://eprint.iacr.org/2020/1281.

Analysing the feasibility of using Objectosphere for Face Presentation Attack Detection*

Lazaro J. Gonzalez-Soler[1] [0000−0001−6470−2966],
Marta Gomez-Barrero[2] [0000−0003−4581−5353],
Manuel Günther[3] [0000−0003−1489−7448], and
Christoph Busch[1] [0000−0002−9159−2923]

[1] dasec - Biometrics and Internet Security Research Group
Hochschule Darmstadt, Germany
{lazaro-janier.gonzalez-soler,christoph.busch}@h-da.de
[2] Hochschule Ansbach, Ansbach, Germany
marta.gomez-barrero@hs-ansbach.de
[3] University of Zurich, Department of Informatics, Switzerland
guenther@ifi.uzh.ch

Abstract. Facial recognition systems have considerably evolved in recent years and have been deployed in a number of real-world applications requiring high security: bank account access, smartphone unlock, and border control, among others. In spite of their advantages, those biometric systems are still vulnerable to attack presentations which can be easily launched to gain access to the aforementioned applications. In order to prevent such threats, several sophisticated Presentation Attack Detection (PAD) techniques have been proposed. These methods aim to detect whether a sample stems from a live subject or from an artificial replica. Such techniques have reported a high detection performance when the attack types or Presentation Attack Instrument (PAI) species are known a priori. However, their accuracy decreases when the test sample's properties remain unknown. In order to enhance the generalisation capability of recent PAD methods based on Convolutional Neural Networks, we explore in this work the feasibility of using Objectosphere for PAD. This loss function has produced reliable results in applications where unknown or innocent subjects have to be rejected while enlisted subjects must be correctly identified (e.g., Watchlist). The experimental evaluation carried out over several challenging scenarios shows that Objectosphere is capable to outperform on average the results attained by the Binary Cross Entropy (BCE) function loss when PAI species are known a priori (a D-EER of 1.50% for Objectosphere vs. 1.71% for BCE).

Keywords: Presentation attack detection · Deep neural network energy · Unknown attacks · Face.

* This research work has been funded by the DFG-ANR RESPECT Project (406880674), and the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

1 Introduction

Face recognition systems have experienced a large development in recent years. They have been successfully applied in numerous unattended applications due to their security, efficiency, and user-friendly data acquisition. In spite of their advantages, face recognition systems can be circumvented by Attack Presentations (APs), in which an artificial representation of the victim's face, denoted as Presentation Attack Instrument (PAI), is presented to a capture device [7]. The extensive evolution achieved by social networks (e.g., LinkedIn, YouTube, or Instagram) allows malicious attackers to download a photo from an authorised subject and re-use it to gain access to several applications. In addition, the creation of sophisticated PAIs such 3D mask [18], make-up [20], and virtual reality [22] is a real threat for face recognition systems.

In order to address such security threats, several Presentation Attack Detection (PAD) approaches have been proposed. They aim to determine whether a sample stems from a live subject (i.e., it is a bona fide presentation - BP) or from an artificial replica (i.e., it is an AP). In this context, most PAD algorithms rely on the limitations of PAIs and their quality degradation during recapture. Based on that fact, hand-crafted PAD methods mainly focused on the analysis of color [1], texture [9, 10], motion [16], and involuntary gestures [15] to distinguish an attack from a bona fide presentation. In addition, the great success of Convolutional Neural Networks (CNNs) has led to the development of powerful architectures which outperform those earlier approaches: those CNN schemes [8, 17, 19] have reported a remarkable detection performance to spot attack types or PAI species when those are included in the training set. However, they suffer a performance degradation when the PAI species remain unknown.

In order to improve the generalisation capabilities of CNNs for PAD, we focus on a different approach known as Objectosphere, which reported a reliable biometric performance for an open-set subject identification scenario in [12]. Since PAD could be seen as an open-set problem where unknown attacks are commonly launched [21], we combine the Objectosphere approach [6, 12] with a CNN based on DenseNet [13] in order to improve the detection of samples stemming from challenging unknown scenarios (e.g., unknown attacks, capture device interoperability, and unknown environmental conditions). The experimental evaluation conducted under the aforementioned scenarios taken from freely available databases such as CASIA Face Antispoofing (CASIA-FASD) [23], REPLAY-ATTACK [3], REPLAY-MOBILE [4], and OULU-NPU [2], reports a PAD performance increase in comparison with the state-of-the-art baselines when PAI species are known a priori (a D-EER of 1.50% for Objectosphere vs. 1.71% for BCE).

The remainder of this paper is organised as follows: the proposed PAD method based on Objectosphere is presented in Sect. 2. The experimental protocol to evaluate the method is explained in Sect. 3. The experimental results benchmarking the performance of our proposal with the top state-of-the-art techniques are discussed in Sect. 4. Finally, conclusions and future work directions are presented in Sect. 5.

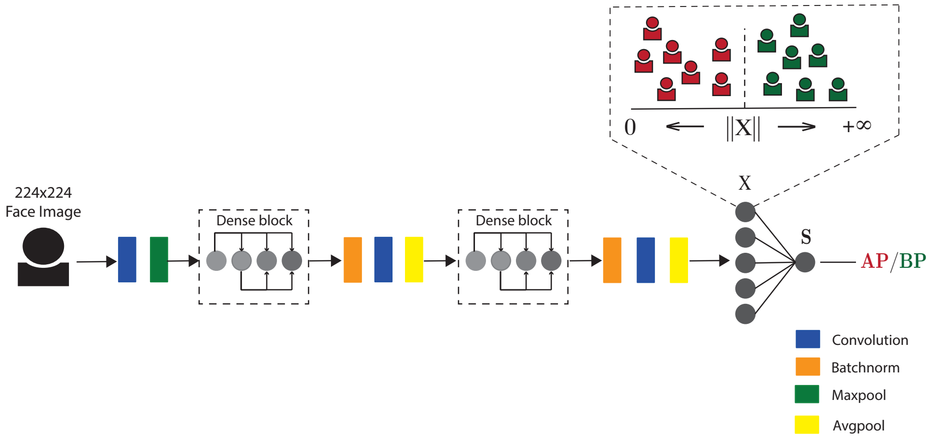


Fig. 1. Overview of our proposed CNN-based algorithm which decreases the energy (i.e., $||X||$) for PAI samples.

2 Proposed Method

Fig. 1 shows an overview of the proposed method based on the Objectosphere. The main goal is to minimise the magnitude or energy of the base network for PAIs whilst increasing the energy for BPs. This way, the network magnitude can be employed as a complementary information for the final decision. In other words, we can easily distinguish both types of presentations by their energy. In our work, DenseNet [13] is used as a base of the final architecture.

2.1 Objectosphere

Objectosphere (Obs) loss was introduced in [6], being an extension of the entropic open-set loss J_E. The latter tries to maximise the entropy for those samples defined as unknown in an open-set scenario with $C > 2$ categories by making their softmax responses $S_c(x)$ uniform. J_E can be computed as:

$$J_E(x) = \begin{cases} -\log S_c(x) & \text{if } x \text{ belong to know classes} \\ -\frac{1}{C}\sum_{c=1}^{C} \log S_c(x) & \text{if } x \text{ is unknown} \end{cases} \tag{1}$$

Building upon the J_E, the Objectosphere loss J_O is computed as:

$$J_O(x) = J_E + \lambda \begin{cases} \max(\xi - \|X\|, 0)^2 & \text{if } x \text{ is known} \\ \|X\| & \text{if } x \text{ is unknown} \end{cases}, \tag{2}$$

where ξ is a constraint applied over the magnitudes of the feature representations, and λ is a regularisation parameter.

Table 1. DenseNet-based architecture used in our experiments.

Layers	Output Size	Filter Size
Convolution	112 × 112	7 × 7, stride 2
Pooling	56 × 56	3 × 3 max pool, stride 2
Dense block (1)	56 × 56	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 6$
Transition block (1)	56 × 56	1 × 1 conv
	28 × 28	2 × 2 average pool, stride 2
Dense block (2)	28 × 28	$\begin{bmatrix}1\times1 \text{ conv}\\ 3\times3 \text{ conv}\end{bmatrix}\times 12$
Transition block (2)	28 × 28	1 × 1 conv
	14 × 14	2 × 2 average pool, stride 2
Classification Layer	X-dimensional fully-connected	
	1-dimensional fully-connected, sigmoid	

Given that the PAD task is considered as a binary classification problem where there exist two labels for the output variable (i.e., BP vs AP), the Objectosphere loss cannot be directly applied to the network's optimisation. Therefore, we combine in our work the second term of J_O with a Binary Cross Entropy (BCE) loss, which is generally employed for binary classification tasks [11]. BCE $J_B(\cdot)$ is computed as:

$$J_B(x) = y \cdot \log S(x) + (1-y) \cdot \log(1 - S(x)), \tag{3}$$

where $S(\cdot)$ is the sigmoid activation function and y is the true label for the input x. We assign $y = 1$ for BPs and $y = 0$ for APs.

Thus, the Objectosphere $J_{O'}$ is reformulated for PAD as:

$$J_{O'}(x) = J_B(x) + \lambda \left[y \cdot \max(\xi - \|X\|, 0)^2 + (1-y) \cdot \|X\| \right] \tag{4}$$

2.2 Network Architecture

As it was mentioned, the architecture used in our investigation is based on the DenseNet model proposed by Huang *et al.* [13]. The network connects each layer to every other layer in a feed-forward fashion as long as they have the same feature map size. This idea reduces the vanishing gradient problem as the dense connections introduce short paths from inputs to outputs [13]. In addition, DenseNet allows implicit deep supervision since the individual layers receive supervision from the loss function due to the shorter paths.

In our work, we combine the first eight layers of DenseNet with a fully connected layer (i.e., X) which, in turn, feeds the decision units S. The selected layers of DenseNet were initialised with the weights trained on the ImageNet database [5]. The eight layers comprise two dense blocks and two transition blocks, as shown in Fig. 1. The dense blocks consist of dense connections between every layer with the same feature map size. The transition blocks normalise and down-sample the feature maps. The network architecture used in our work is summarised in Tab. 1.

3 Experimental Protocol

The experimental evaluation aims to: *i*) analyse the impact of Objectosphere in terms of detection performance for several CNN topologies, *ii*) study the Objectosphere detection performance for challenging unknown attacks, and *iii*) establish a benchmark with the state-of-the-art PAD techniques. In order to reach our goals, we focus on three different scenarios:

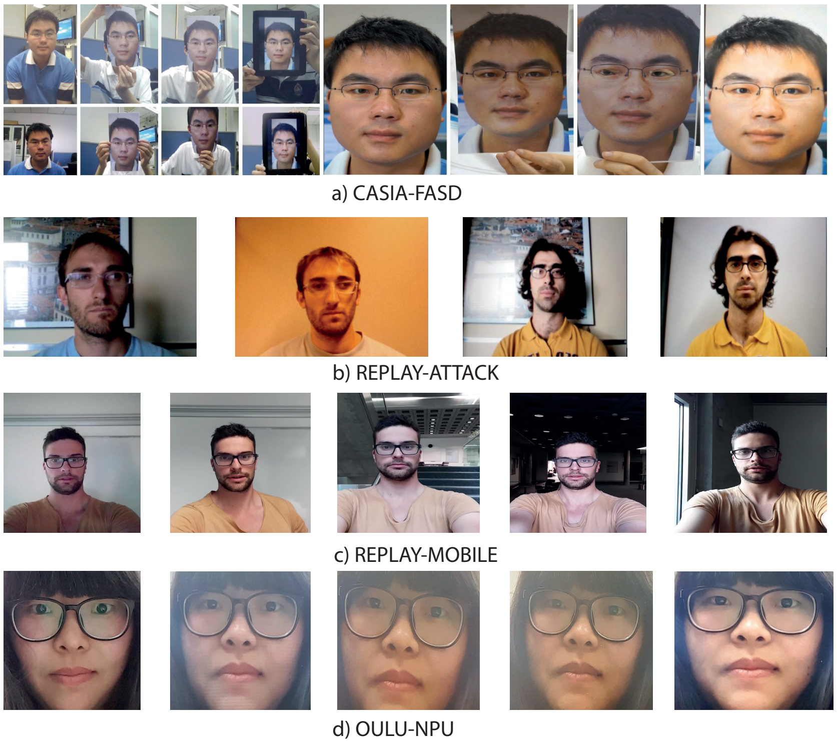
- *Known-attacks*, which includes an analysis of all PAI species. In all cases, PAI species for testing are also included in the training set, as described in [23]. In this scenario, we analyse the impact of Objectosphere in terms of detection performance for several size of the embedding X.
- *Unknown PAI species*, in which the PAI species used for testing are not incorporated in the training set. We use the protocols described in [2].
- *Cross-database*, in which the datasets employed for testing are different from the databases used for training. Both datasets contain the same PAI species to ensure that the performance degradation is due to the dataset change and not to the unknown PAI species.

3.1 Databases

The experimental evaluation is conducted over four well-established databases: CASIA-FASD [23], REPLAY-ATTACK [3], REPLAY-MOBILE [4], and OULU-NPU [2]. Fig. 2 shows samples for each database used.

CASIA-FASD [23] contains 600 short videos of bona fide and attack presentations stemming from 50 different subjects and acquired under different conditions. The dataset comprises three PAI species: *i*) warped photo attacks or printed attacks, in which the attackers place their face behind the hard copies of high-resolution digital photographs, *ii*) cut photo attacks, the face of the attacker is placed behind the hard copies of photos, where eyes have been cut out, and *iii*) video replay attacks, where attackers replay face videos using iPads.

REPLAY-ATTACK (RA) [3] consists of 1200 short videos (around 10 seconds in mov format) of both bona fide and attack presentations of 50 different subjects, acquired with a 320 × 240 low-resolution webcam of a 13-inch MacBook Laptop. The video samples were recorded under two different conditions: *i*) controlled, with uniform background and artificial lighting, and *ii*) adverse, with natural illumination and non-uniform background. In addition, this database comprises three PAI species: printed attacks, photo replay attacks (i.e., a photo is replayed by a smartphone to the capture device), and video replay attacks.


a) CASIA-FASD

b) REPLAY-ATTACK

c) REPLAY-MOBILE

d) OULU-NPU

Fig. 2. Examples of images include in the databases used in this work.

REPLAY-MOBILE (RM) [4] comprises 1190 video clips of printed attacks, photo replay attacks, and video replay attacks of 40 subjects under different lighting conditions. Those videos were recorded with two smart capture devices: an iPad Mini2 and a LG-G4 smartphone, thereby allowing the evaluation of PAD approaches for the mobile scenario.

OULU-NPU [2] consists of 4950 high-resolution short video sequences of BPs and AP attempts stemming from 55 subjects. The BP samples were captured in three different sessions with different illumination conditions and background scenes. The PAI species are printed attacks and video replay attacks which were recorded using the frontal cameras of six mobile phones. This database defines four different protocols which are considered for the unknown attack evaluation:

- Protocol 1 evaluates the generalisation capability of PAD techniques under different environment conditions (i.e., illumination and background scenes).
- Protocol 2 is designed to evaluate the PAD generalisation capability when tested PAI species remain unknown from the training set.
- Protocol 3 analyses the capture device interoperability. To that end, a Leave One Camera Out (LOCO) is followed. Thus, for each iteration, samples

recorded by five smartphones are used for training whilst images captured by the sixth mobile device is used in the evaluation.

- Protocol 4 is the most challenging scenario as it combines all described protocols. Specifically, the generalisation capability of PAD methods is simultaneously evaluated across previously unknown illumination conditions, background scenes, PAI species, and capture devices.

3.2 Evaluation Metrics

Finally, all results are analysed and reported in compliance with the metrics defined in the international standard ISO/IEC 30107-3 [14] for biometric PAD:

- Attack Presentation Classification Error Rate (APCER), which is defined as the proportion of attack presentations wrongly classified as bona fide presentations.
- Bona Fide Presentation Classification Error Rate (BPCER), which is the proportion of bona fide presentations misclassified as attack presentations.

Based on the above metrics, we report *i*) the Detection Error Trade-off (DET) curves between APCER and BPCER; *ii*) the BPCERs observed at different APCER values or security thresholds such as 10% (BPCER10), 5% (BPCER20), and 1% (BPCER100), respectively; and *iii*) the Detection Equal Error Rate (D-EER), which is defined as the error rate value at the operating point where APCER = BPCER.

4 Experimental Results

4.1 Impact on Network Topology

In the first set of experiments, we explore the impact of Objectosphere on different size of the embedding X. To that end, we compute in Fig. 3 the error rates of our proposal for different numbers of units in $X = \{16, 32, 64, 128, 256, 512\}$ over CASIA, RA, and RM databases. We follow the know-attack scenario defined in Sect. 3. The parameters $\xi = 20$ and $\lambda = 10^{-3}$ involved in the Objectosphere optimisation were selected from [6]. We can observe that the training of the CNN using BCE reports different errors rates at different numbers of neural units (see Fig. 3-a), thereby yielding a D-EER of 1.71% ± 1.64. In contrast, its optimisation through Objectosphere appears to be more stable across the number of neural units, thereby resulting in a D-EER of 1.50% ± 1.78. These results can be clearly perceived in Fig. 3-b). It is important to highlight that the evaluation over RM yields a D-EER of 0.0% for all neural units optimised separately by both BCE and Objectosphere. Hence, the results are not depicted in Fig. 3.

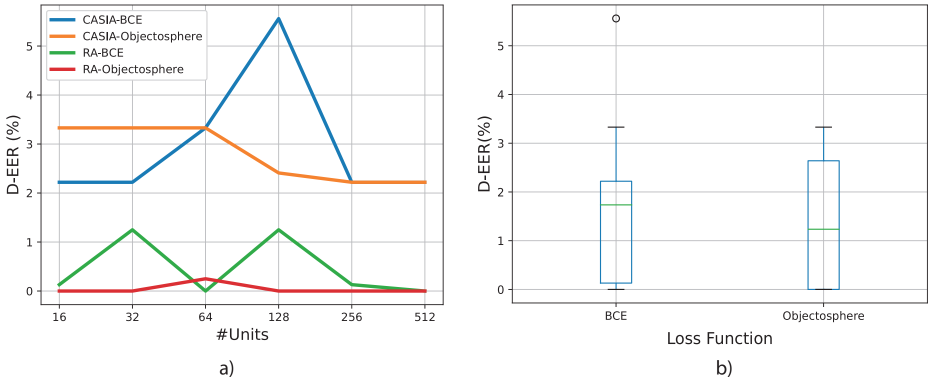


Fig. 3. Impact of loss function over the CNN topology. a) Detection performance of loss function per neural unit for CASIA and RA, and b) detection performance per loss function.

Table 2. Benchmark in terms of D-EER(%) of our proposed method for the number of units $X = 512$ over different protocols in OULU-NPU.

Method	**Protocol 1**	**Protocol 2**	**Protocol 3**	**Protocol 4**
FAS-BAS [17]	1.60	**2.70**	**2.70 ± 1.30**	**9.30 ± 5.60**
IQM-SVM [8]	19.17	12.50	21.94 ± 9.99	34.17 ± 25.84
LBP-SVM [8]	12.92	30.00	28.50 ± 23.05	41.67 ± 27.03
DeepPixelBis [8]	**0.83**	11.39	11.67 ± 19.57	36.67 ± 29.67
BSIF-FV [10]	7.81	8.33	7.12 ± 2.56	11.04 ± 2.00
Proposal (BCE)	3.33	4.24	6.50 ± 3.68	13.90 ± 8.37
Proposal (Objectosphere)	3.33	4.17	3.41 ± 3.53	15.23 ± 12.21

4.2 Challenging Unknown Attacks

Now, we select the number of units where both BCE and Objectosphere reported best detection performance (i.e., $X = 512$) and compute in Tab. 2 their D-EER over challenging OULU-NPU protocols [2]. As it can be observed, the proposed method optimised through Objectosphere reports a detection performance improvement in comparison with the network trained with BCE. In particular, the former achieves for the protocol 3 a D-EER of 3.41% ± 3.53 which is almost twice lower than the one attained by the proposal with BCE (i.e., 6.50 ± 3.68). In addition, both Objectosphere and BCE are capable of achieving a state-of-the-art detection performance under most of the protocols. Especially for the challenging protocols 2, 3, and 4 which evaluate unknown PAI species, capture devices, and environmental conditions, respectively.

On the other hand, we carry out a depth analysis of the detection performance of our proposed algorithm and establish a benchmark between BCE and Objectosphere for several system operating points in Fig. 4. As it can be ob-

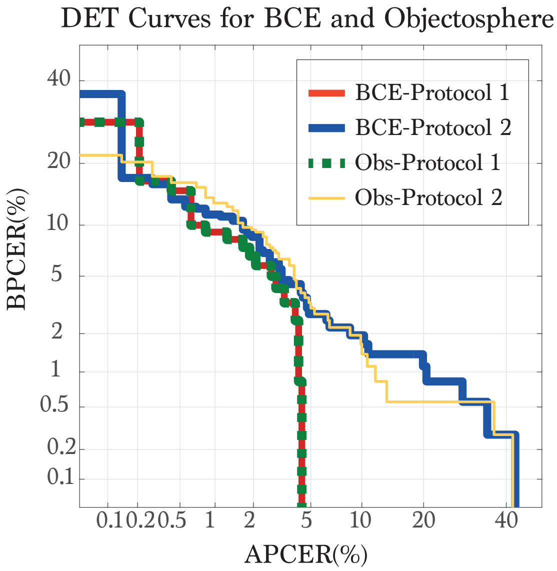


Fig. 4. Benchmark of Objectosphere with BCE for different operating points for unknown environment conditions and PAI species.

served, Objectosphere reports the best detection performance for higher operating points. In particular, it achieves, for Protocol 2, a BPCER in around 22.00% for an APCER of 0.0% which is almost twice lower than the one yielded by the BCE. In addition, the curves for both optimisers show a similar behaviour for Protocol 1, thereby resulting in a BPCER of 0.0% for any APCER $\geq$ 5.0%.

4.3 Cross Database

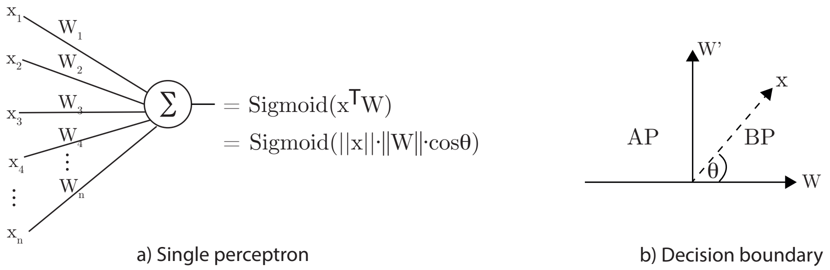
Finally, we explore the capability of Objectosphere to improve the PAD performance across different databases which include images with varying resolutions, lighting conditions, subjects, etc. Such databases are unknown in training time, thereby leading to a challenging cross-database evaluation [10]. To that end, we select the range of $X = \{16, 32, 64, 128, 256, 512\}$ defined above and report the corresponding D-EERs in Tab. 3. As it can be seen, the CNN optimised through Objectosphere achieves similar results to those reported by BCE. In addition, a high variability of results can be perceived for different numbers of units, especially for RA which contains low-quality images.

4.4 Performance Discussion

As it can be appreciated from the above results, Objectosphere allows improving the PAD in most cases. However, it still shows a limited capability to generalise to those scenarios whose images pose characteristics different to the ones used

Table 3. Benchmark in terms of D-EER(%) of our proposed method for several number of units over the cross-database evaluation.

Train	Test	Method	16	32	64	128	256	512	Avg.
CASIA	RA	BCE	20.00	30.00	36.25	31.25	28.75	27.50	28.96
		Obs	21.75	28.75	25.00	30.00	23.75	28.75	**26.33**
	RM	BCE	8.52	12.87	9.49	17.23	12.87	13.59	**12.43**
		Obs	10.21	14.56	9.48	15.80	10.92	14.30	12.55
RA	CASIA	BCE	48.89	52.59	53.33	53.33	50.19	47.78	51.02
		Obs	47.96	48.89	50.00	50.19	47.78	45.74	**48.43**
	RM	BCE	63.59	66.30	60.66	71.85	68.21	64.30	65.82
		Obs	58.97	60.21	50.00	56.57	52.67	43.43	**53.64**
RM	CASIA	BCE	26.86	23.33	28.89	30.00	30.00	30.00	**28.18**
		Obs	25.56	33.52	30.00	27.78	31.30	35.56	30.62
	RA	BCE	21.38	23.75	23.00	21.25	20.25	25.50	**22.52**
		Obs	27.75	19.13	26.25	20.00	22.50	27.50	23.86



a) Single perceptron

b) Decision boundary

Fig. 5. Math operation performed between an input vector x and learned weights W. a) Single perceptron with **n** units and a sigmoid activation function and b) two-dimensional example of a decision boundary for the sigmoid function.

for training the CNN. In order to extend the expandability of our results, we perform an ablation study in which mathematical operations are analysed on a single perceptron tuned with Objectosphere. Fig. 5 shows the basic operation performed on a single perceptron. This perceptron performs a dot product between an input vector (i.e., x) and the weights (i.e., W) learned on a feed-forward fashion[4]. This dot product can be seen as the angle between x and W. Therefore, it can be computed as $||x|| \cdot ||W|| \cdot \cos\theta$, which is represented in Fig. 5-b). In addition, we note that the BP vs. AP decision boundary for the sigmoid function is determined by W' an orthogonal vector to W, i.e., all those solutions in the second quadrant in Fig. 5-b) are regarded as an AP while solutions in the first quadrant are taken as a BP.

[4] We omit the bias in the single perceptron operations, as it is ignored in our implementation.

Keeping those considerations in mind, we observe that increasing or decreasing either the length, magnitude, or energy of x or W would change no solutions as the AP vs. BP decision is conditioned by the angle θ between both vectors: $\theta \geq 90°$ is considered the input sample as an AP. Otherwise, the facial image at hand is a BP. Based on that fact, we strongly think that future directions should be focused on angle optimisation instead of vector magnitude to improve the BCE results. In addition, it could also be combined with Objectosphere for further improvements.

5 Conclusion

In this work, we evaluate the feasibility of using Objectosphere for facial PAD. Objectosphere is a loss function developed for open-set categorical cross-entropy which tries to enhance the generalisation capabilities of CNNs for classification problems with more than two categories. Since PAD is considered a binary classification problem, we extended Objectosphere to be used in combination with Binary Cross Entropy. The experimental results conducted in compliance with the international ISO/IEC 30107-3 [14] and over several challenging and realistic scenarios reported a slight improvement of Objectosphere with respect to the baseline BCE. Specifically, the former yielded a D-EER of 1.50% which is lower than the one attained by BCE (i.e., D-EER of 1.71%) for known-attack scenarios.

In spite of those positive results, we noted that Objectosphere has a limited generalisation capability for those scenarios where either the PAI species or capture devices are not included in the training set, thereby resulting in a similar performance with respect to the baseline (i.e., BCE). In order to extend the expandability of our results, we carried out a performance study on basic math operations performed at neural unit level. We showed that the sigmoid decision boundary is determined by the angle between the input vector and the weights learned by the CNN. As future directions, we plan the development of new angle-based strategies to enhance the BP vs. AP decision in a binary classification task.

References

1. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: Proc. Intl. Conf. on Image Processing (ICIP). pp. 2636–2640 (2015)
2. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: Proc. Intl. Conf. on Automatic Face & Gesture Recognition (FG). pp. 612–618 (2017)
3. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proc. Intl. Conf. on Biometrics Special Interests Group (BIOSIG) (2012)

4. Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., Marcel, S.: The replay-mobile face presentation-attack database. In: Proc. Intl. Conf. on Biometrics Special Interests Group (BIOSIG) (2016)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
6. Dhamija, A.R., Günther, M., Boult, T.E.: Reducing network agnostophobia. In: Proc. Intl. Conf. on Neural Information Processing Systems (NIPS) (2018)
7. Galbally, J., Marcel, S., Fierrez, J.: Biometric antispoofing methods: A survey in face recognition. IEEE Access **2**, 1530–1552 (2014)
8. George, A., Marcel, S.: Deep pixelwise binary supervision for face presentation attack detection. In: Proc. Intl. Conf. on Biometrics (ICB). pp. 1–8 (2019)
9. Gonzalez-Soler, L.J., Gomez-Barrero, M., Busch, C.: Fisher vector encoding of dense-bsif features for unknown face presentation attack detection. In: Proc. Intl. Conf. of the Biometrics Special Interest Group (BIOSIG). pp. 1–6. IEEE (2020)
10. Gonzalez-Soler, L.J., Gomez-Barrero, M., Busch, C.: On the generalisation capabilities of fisher vector based face presentation attack detection. IET Biometrics pp. 1–12 (2021)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
12. Günther, M., Dhamija, A.R., Boult, T.E.: Watchlist adaptation: Protecting the innocent. In: Proc. Intl. Conf. of the Biometrics Special Interest Group (BIOSIG). pp. 1–7 (2020)
13. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. Intl. Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4700–4708 (2017)
14. ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. International Organization for Standardization (2017)
15. Jee, H.K., Jung, S.U., Yoo, J.H.: Liveness detection for embedded face recognition system. Proc. Intl. Journal of Biological and Medical Sciences **1**(4), 235–238 (2006)
16. Kollreider, K., Fronthaler, H., Bigun, J.: Non-intrusive liveness detection by face images. Image and Vision Computing **27**(3), 233–244 (2009)
17. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 389–398 (2018)
18. Manjani, I., Tariyal, S., Vatsa, M., Singh, R., Majumdar, A.: Detecting silicone mask-based presentation attack via deep dictionary learning. IEEE Trans. on Information Forensics and Security **12**(7), 1713–1723 (2017)
19. Muhammad, U., Hadid, A.: Face anti-spoofing using hybrid residual learning framework. In: Proc. Intl. Conf. on Biometrics (ICB) (2019)
20. Rathgeb, C., Drozdowski, P., Busch, C.: Detection of makeup presentation attacks based on deep face representations. In: Proc. Int. Conf. on Pattern Recognition (ICPR). pp. 1–6. IEEE (April 2020)
21. Xiong, F., AbdAlmageed, W.: Unknown presentation attack detection with face rgb images. In: Proc. Intl. Conf. on Biometrics Theory, Applications and Systems (BTAS). pp. 1–9 (2018)
22. Xu, Y., Price, T., Frahm, J., Monrose, F.: Virtual u: Defeating face liveness detection by building virtual models from your public photos. In: Proc. USENIX. pp. 497–512 (2016)
23. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: Proc. Intl. Conf. on Biometrics (ICB). pp. 26–31 (2012)

Enhancing FIDO Transaction Confirmation with Structured Data Formats

Andre Büttner[0000−0002−0138−366X] and Nils Gruschka[0000−0001−7360−8314]

University of Oslo, Gaustadalléen 23B, 0373 Oslo, Norway
{andrbut,nilsgrus}@ifi.uio.no

Abstract. FIDO Transaction Confirmation is an extension for the FIDO authentication protocols to enable the verification and signing of digital transactions, e.g., for online banking. The standard currently considers only to include a transaction message text in the assertion which is signed by the user's authenticator. However, this is not useful for more complex transactions and leaves room for ambiguities that might lead to security vulnerabilities. Therefore, we propose to include the transaction information to the FIDO protocols in a structured data format with a strictly defined schema to validate and sign transactions more reliably and securely.

Keywords: FIDO · Transactions · Security

1 Introduction

In recent years, passwords have proven to be not secure enough to withstand attacks, such as, phishing or brute-forcing [3]. Consequently, two-factor and multi-factor authentication have been introduced to make authentication more secure [1]. The FIDO Alliance has proposed protocols for using authenticators as an additional factor and even as a passwordless solution. An important extension to these protocols is the *Transaction Confirmation* [2], which allows users to confirm online transactions using a FIDO authenticator. A relying party can include a transaction message or an image to an assertion request, which is displayed to the user and signed by the authenticator. However, research has shown that it is possible to trick a user into approving a malicious transaction [9,10].

Further, since the transaction is only represented as a text string or an image without clear defined semantics, the transaction information leaves room for ambiguities. Therefore the desirable *What-You-See-Is-What-You-Sign* [7] property is not sufficiently fulfilled. It would be more reliable to use a structured data format that contains a well-formed and self-describing representation of a transaction [4,6]. The contribution of this paper is therefore a proposal and discussion on the use of structured data formats for FIDO Transaction Confirmation.

The remainder of this paper is structured as follows. In Section 2 some background on FIDO Transaction Confirmation is provided. Our proposed enhancement for the FIDO transaction extension is described in Section 3. Section 4 discusses advantages and disadvantages of our approach. Finally, in Section 5 our findings are concluded and suggestions on future work are given.

2 FIDO Transaction Confirmation

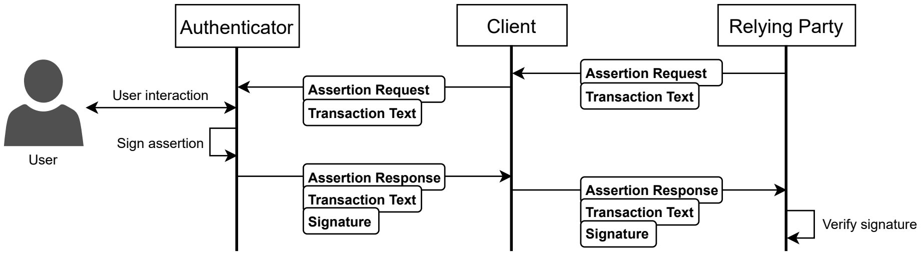


Fig. 1. Transaction Confirmation flow diagram showing the different processing steps.

The FIDO UAF and FIDO2/WebAuthn protocols are based on a challenge-response protocol, where an authenticator, e.g., a smartphone, hardware token or platform authenticator, registers with a public-key against a relying party. For authentication, the authenticator needs to sign a random challenge to proof possession of the corresponding private key.

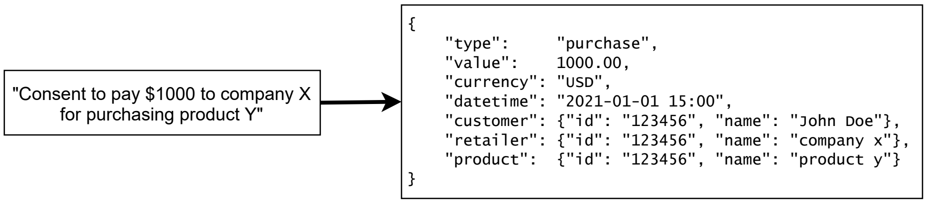
Transaction Confirmation as an extension of these protocols seeks for “a standardized and secure way of gathering explicit user consent for a specific action” [2]. Consent is based on the user’s interaction with the respective authenticator to confirm that he has seen and approved the transaction message. This allows to use FIDO authenticators for carrying out bank transactions, online purchases, granting access to certain information, and more.

Fig. 1 gives an overview on how a transaction is processed with the FIDO protocols. The relying party sends a FIDO assertion request to the client, which contains a human-readable representation of a transaction in form of a simple text. The user confirms the transaction by interacting with the authenticator. Afterwards the authenticator creates the assertion response along with the signature created with the corresponding private key. The assertion response is then returned via the client application to the relying party, which finally verifies the signature and executes the requested transaction [5].

3 Structured Data for Transactions

Instead of just plain text, we propose to use a machine-readable representation of a transaction that is converted into a human-readable text by the client or authenticator. One common data format for structuring data is the Extensible Markup Language (XML), which is typically defined and validated using the XML Schema Definition (XSD) language. Also, there are respective W3C standards for signature generation and encryption.

The data formats used in the FIDO protocols are JavaScript Object Notation (JSON) and its binary counterpart Concise Binary Object Representation



"Consent to pay $1000 to company X for purchasing product Y"

```
{
    "type":     "purchase",
    "value":    1000.00,
    "currency": "USD",
    "datetime": "2021-01-01 15:00",
    "customer": {"id": "123456", "name": "John Doe"},
    "retailer": {"id": "123456", "name": "company x"},
    "product":  {"id": "123456", "name": "product y"}
}
```

Fig. 2. Example transaction as plain text and structured data.

(CBOR). FIDO extensions are expected to be in the CBOR format. Thus, we consider this data format to be most suitable for transaction data structures as well. Similar to XSD, there already exists the Concise Data Definition Language (CDDL) which analogously enables the definition and validation of CBOR objects. Signatures, message authentication and encryption are standardized in the CBOR Object Signing and Encryption (COSE) protocol. In Fig. 2 on the left hand-side, an example mentioned in [2] is shown. With our approach this can be replaced by a semi-structured representation like presented on the right-hand side. Some information like identifiers and time were added, showing how transactions could easily be extended with relevant information. Also, validation and limitations on each of the attributes could be applied by the authenticator. Further aspects are discussed in the following section.

4 Discussion

Semi-structured data formats like XML, JSON or CBOR provide properties, e.g., well-formedness and being self-describing with clear semantics [8]. This avoids ambiguities from unclear formulations, which is common for plain text. Further, for more complex types of transactions it might be useful to display only relevant parts to the user before signing. This can be realized more easily with structured data, if these parts are separate attributes inside the data structure, e.g., an account number inside a bank transaction. Also, structured data is machine-readable, which allows to define policies for certain attributes. These can be provided as CDDL schemas by the relying party during registration, which are then used by the client application or the authenticator for validation.

FIDO transactions may be manipulated or eavesdropped through XSS or malware on the client. Therefore it is reasonable to let the relying party sign [9] and encrypt the transaction data. If the CBOR format is used for transactions, the COSE protocol can provide a standardized way for ensuring both integrity and confidentiality on both ends.

An obvious disadvantage of using data structures for FIDO transactions are the complexity and its data overhead. This may especially be problematic for hardware tokens with limited computation and storage resources. The authenticator would need to perform CDDL schema validation. Ideally, it should also support the COSE signature validation and decryption. Increased latency may be acceptable, since registration and normal authentication would not be affected. In case it does not work on hardware tokens, the validation can be outsourced to the client application, however, reducing the security gain.

5 Conclusion and Future Work

The FIDO protocols are a promising step towards more secure authentication and a potential replacement for passwords. Transaction Confirmation is a good example of how these protocols support use cases beyond that. This paper addresses some shortcomings of this extension and proposes to use structured data instead of plain text. As discussed, our approach provides many opportunities, such as allowing an authenticator to validate transactions against policies and using standardized ways to ensure integrity and confidentiality.

In future work, we are planning to test the approach for different applications and different types of authenticators, to analyze different attack scenarios and to evaluate the application of CDDL schemas and COSE signature and encryption.

References

1. Dasgupta, D., Roy, A., Nag, A.: Multi-Factor Authentication, pp. 185–233. Springer International Publishing, Cham (2017)
2. FIDO Alliance: FIDO Transaction Confirmation White Paper. Tech. rep. (August 2020), https://media.fidoalliance.org/wp-content/uploads/2020/08/FIDO-Alliance-Transaction-Confirmation-White-Paper-08-18-DM.pdf
3. Florêncio, D., Herley, C., Coskun, B.: Do strong web passwords accomplish anything? HotSec **7**(6), 159 (2007)
4. Gruschka, N., Reuter, F., Luttenberger, N.: Checking and signing xml documents on java smart cards. In: Quisquater, J.J., Paradinas, P., Deswarte, Y., El Kalam, A.A. (eds.) Smart Card Research and Advanced Applications VI. pp. 287–302. Springer US, Boston, MA (2004)
5. Hodges, J., Czeskis, A., Liao, H., Lindemann, R., Balfanz, D., Jones, J., Lundberg, E., Kumar, A., Jones, M.: Web authentication: An API for accessing public key credentials level 1. W3C recommendation, W3C (Mar 2019), https://www.w3.org/TR/2019/REC-webauthn-1-20190304/
6. Jøsang, A., AlFayyadh, B.: Robust wysiwys: A method for ensuring that what you see is what you sign. In: Proceedings of the Sixth Australasian Conference on Information Security - Volume 81. p. 53–58. AISC '08, Australian Computer Society, Inc., AUS (2008)
7. Landrock, P., Pedersen, T.: WYSIWYS?—What you see is what you sign? Information Security Technical Report **3**(2), 55–61 (1998)
8. Nenadi, A., Zhang, N.: Non-repudiation and fairness in electronic data exchange. In: Enterprise Information Systems V, pp. 286–293. Springer (2004)
9. Xu, P., Sun, R., Wang, W., Chen, T., Zheng, Y., Jin, H.: SDD: A trusted display of FIDO2 transaction confirmation without trusted execution environment. Future Generation Computer Systems **125**, 32–40 (2021)
10. Zhang, Y., Wang, X., Zhao, Z., Li, H.: Secure display for FIDO transaction confirmation. In: Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy. pp. 155–157 (2018)

Cyber-Physical Tracking of IoT devices: A maritime use case

Ahmed Amro[0000−0002−3390−0772]

Norwegian University of Science and Technology, Gjøvik, Norway
ahmed.amro@ntnu.no

Abstract. We live in a highly connected world. Many types of devices involved in numerous applications are connected to the internet and their number is increasing day by day. In the maritime domain, maritime entities utilize the Internet to connect geographically dispersed vessels, offshore units, and other sorts of components in the maritime infrastructure. While the locations of some of these components are publicly available through various resources (e.g. MarineTraffic), their cyber-related information is not necessarily intended to be. Obtaining the knowledge of both the physical location as well as the cyber-related information of certain components might provide attackers with opportunities to perform more sophisticated and targeted attacks. With new regulations and guidelines aiming to improve cybersecurity in maritime, investigating possible threats against the maritime infrastructure is required. To this end, this paper investigates the issue of combined cyber and physical tracking of IoT devices with a prime focus on maritime infrastructure. I propose a process for Cyber-Physical tracking of IoT devices including maritime components that are connected to the internet. I employed several IoT scanners (e.g. Shodan) and obtained cyber-related information as well as their physical properties such as location, speed, and others. I have identified 4942 hosts that emit NMEA messages and 331 possible maritime components. Furthermore, I provide discussion regarding the expected risks of such a process while considering both the current state of affairs in maritime as well as futuristic operational modes such as autonomous, unmanned, and remotely connected vessels.

Keywords: IoT · scanning · tracking · cybersecurity · maritime · NMEA

1 Introduction

The amount of connected devices to the internet is growing each year and is expected to double from the year 2021 to 2025 reaching 75.44 billion [28]. This rapid trend of connectivity can pave the way for innovation and an improved way of life, however, it introduces a wide range of cyber attacks if cybersecurity is not considered during the development of such devices and their hosting systems.

Several sectors are following the trend of Internet of Things (IoT) and Industrial IoT (IIoT) including maritime [15, 5]. The maritime sector is undergoing a digital transformation era that drastically impacts its technologies, business models and operations [10]. Vessel tracking services are among these operations as maritime operators must keep track of their vessels and geographically distinct components for improved management. Therefore, they rely on devices that are connected to the internet and emit marine information that is important for their operations such as location, speed, heading, and others. Some of these devices employ a protocol proposed by National Marine Electronics Association (NMEA) for communicating among maritime components. Under normal circumstances, vessel tracking is a very common domain and field of study. Some marine traffic information is a publicly open resource utilized for legitimate ship tracking purposes. However, fingerprinting and cyber tracking of vessels by unauthorized entities is a less-discussed subject. Previously, ships have been fingerprinted and tracked using data from Automatic Identification Systems (AIS) [24]. Such activities can be conducted by attackers during the reconnaissance stage toward the development of more advanced and targeted attacks. Attackers can collect cyber-related information about target ships such as their connected devices and their vulnerabilities and use this information during exploitation (further discussed in section 4).

Recently, the International Maritime Organization (IMO) has passed Resolution MSC. 428(98) [6] for maritime risk management. The resolution makes it mandatory for ship owners and operators to include cybersecurity in their safety management systems. Among the discussed risk management activities in the resolution is continuous risk analysis considering the threat landscape. My paper supports the efforts in this direction by capturing the current state of a very common maritime protocol that is NMEA observed on the internet. I follow a state-of-the-are process for IoT vulnerability scanning proposed in my earlier work [1] and utilize known IoT scanners (e.g. Shodan) for detecting NMEA emitting devices. Moreover, I develop upon the approach of detecting vessels using AIS data and utilize NEMA messages for fingerprinting maritime components using them. My work aims to shed the light on a possible threat against organizations and systems employing NMEA. My contributions in this paper are summarized as follows:

- I propose a process for Cyber-Physical tracking of IoT devices with a prime focus on maritime components. This process emulates an adversarial behavior against systems using NMEA as an early stage of cyber attacks.
- I present the current status of NMEA service considering the type of messages, devices, ports, and countries. I believe that this information is valuable for the cybersecurity community in maritime and other sectors employing NMEA.
- I provide discussion regarding the risks of my proposed Cyber-Physical tracking process considering both the current state of affairs in the maritime domain as well as considering futuristic operational modes.

The remainder of this paper is organized as follows. Section 2 discusses relevant concepts and artifacts that are utilized in this paper. Then, section 3 discusses in detail my proposed Cyber-Physical tracking process which also resulted in capturing the status of NMEA messages on the internet. Afterward, section 4 provides a discussion of the risks associated with my proposed process, provides suggestions for mitigation, and discusses limitations. Finally, section 5 concludes the work in this paper.

2 Background and Related Work

Shodan [16]; a known IoT search engine has previously presented a ship tracking capability utilizing AIS data communicated over the Internet which includes position information. This has been argued to be a wake-up call for maritime cybersecurity [24]. Since then, very limited works have discussed this issue as the number of AIS-connected devices visible to the internet are very limited according to my latest search for AIS messages (e.g AIVDM, AIVDO, ABVDM, etc) on Shodan. That work highlighted the ability for unauthorized entities to gain both physical and cyber information regarding vessels by relying on protocols that are accessible through cyber means and disclose physical properties. That work had led me to consider NMEA protocol as another approach. I believe that NMEA provides a suitable link between both the cyber and physical realms.

There are several NMEA protocols including NMEA0183 [2] and NMEA2000 [14]. NMEA messages abiding by the NMEA0183 protocol are textual messages containing structured information intended originally for navigation purposes. The format of NMEA messages includes static information and dynamic information. The static information includes a TalkerID and a MessageID. The dynamic information includes several fields each containing specific information such as time, longitude, latitude, and others (refer to [2, 26] for more details). This information is utilized in legitimate vessel tracking services as well as legitimate navigational functions. The messages are not encrypted or encoded, they are communicated in plain text. Therefore, they can be used to fingerprint devices emitting them.

Originally, NMEA0183 are mostly transmitted over serial links restricting access to them to specific systems and locations [2]. However, adaptations have been proposed to transmit NMEA0183 messages over TCP and UDP protocols making them accessible through IP networks. This transformation introduced a wide range of cyber attacks. The security; or in better terms, the lack of security in NMEA has been discussed by several works. Tran et al [31] have discussed the security of several marine protocols including NMEA0183. The authors referred to the lack of authentication, encryption, and validation of NMEA messages. The authors argued that the messages are susceptible to many attacks if attackers can identify the network device that uses the standard. Other works have argued that NMEA security currently depends on the network and host security [27, 9].

My research targets maritime risk management with a current focus on the risks related to NMEA messages. I employ the *ATT&CK* framework [29] for threat modeling to identify threats against maritime systems and components across the different adversarial tactics (i.e. kill chain phases) of cyber attacks. This paper considers attack techniques and mitigation related to NMEA messages during the reconnaissance stage of cyber attacks. I investigate and demonstrate the ability of attackers to fingerprint devices emitting NMEA messages over the internet. Future work will focus on subsequent kill chain phases.

In this paper, I rely on my previous work [1] in which I presented the state-of-the-art in IoT scanning and vulnerability scanning. I highlighted the increased interest in the field, discussed some challenges, and proposed a systematic process for IoT vulnerability scanning. I also proposed a scanning space in which all scanning processes occur. The space consists of three dimensions, namely, IP addresses, ports, and vulnerabilities. The IP addresses specify the range of hosts to scan for, the ports specify the range of services to connect to, while the vulnerabilities specify which type of vulnerabilities the scan process is looking for. I referred to the Open Web Application Security Project (OWASP) which published the top 10 IoT vulnerability categories [17]. Among the discussed vulnerabilities is insecure data transmission and storage which is relevant to NMEA as the messages are transmitted in plain text and are susceptible to a wide range of attacks. In this paper, I propose a new attack technique by exploiting the insecure manner in which NMEA messages are transmitted and using them in fingerprinting specific targets and gathering victim information for targeted attacks.

3 Cyber-Physical Tracking of Maritime Components

In this section, I describe my proposed methodology for cyber-physical tracking of maritime components. An overview of my approach is depicted in Figure 1.

My hypothesis is that some maritime components such as vessels have both cyber and physical properties. Cyber properties, include; among others, IP address, services, data, and vulnerabilities. These properties can be recorded by Internet-wide scanners such as Shodan if they are publicly communicated through the Internet. The data might include maritime-specific protocols such as NMEA which I propose to be used to fingerprint maritime components. On the other hand, the physical properties include; among others, navigation information such as location, speed, heading, and time of fix. Some marine tracking services such as MarineTraffic receive such information from various sources, record them and make them available for the public. My approach for correlating these two resources towards the identification and tracking of maritime components is guided by the state-of-the-art process of IoT scanning presented in my earlier work [1]. The process starts with selecting suitable scanner tools and configuring them with the suitable parameters to achieve the objective of the scanning process. Then, the scanning process is initiated and the results are collected. Afterward, the results are validated and analyzed. A detailed description of each step is discussed hereafter.

3.1 Scanner Selection

There are many networks and IoT scanners discussed in the literature. However, Shodan and Censys are the most referenced as stated in my earlier work [1]. The

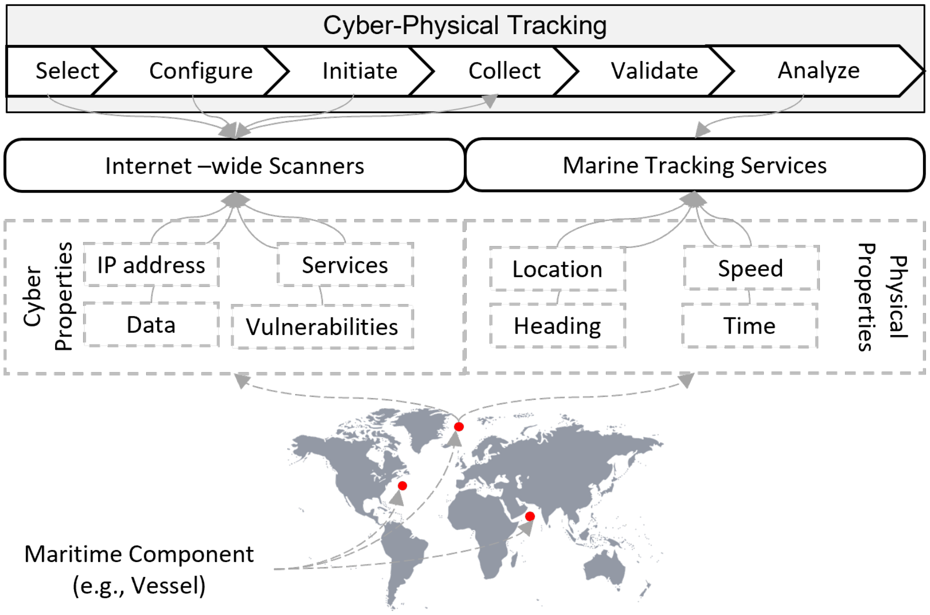


Fig. 1. Overview of my methodology for Cyber-Physical Tracking of maritime components

same notion is observed in several works in the literature. Li et al[13] have presented a survey of Internet-wide scanners including Shodan, Censys, ZoomEye, BinaryEdge, and Fofa. Another review of several public network vulnerability scanners is presented by Tundis et al [32]. The authors have discussed and evaluated Shodan, Censys, ZoomEye, Thingful, and PunkSpider. Moreover, Fofa and BinaryEdge scanners have been utilized in state-supported cyber activities as referred to in the report regarding Irans secret cyber files recently this year [11].

I have previously discussed two types of scanning approaches implemented in the different scanner tools, namely, passive scanning and active scanning [1]. Active scanning is the act of actively attempting to initiate connections with devices in a certain scope that can include the entire internet and recording their responses. On the other hand, passive scanning is the act of querying an indexed database that stores results of previous active scanning activities [1]. In this paper, I have followed the passive scanning approach to avoid any access violations. Therefore, I considered the most common scanners, namely, Shodan [16] and Censys [8], all of which allow for passive scanning while they are conducting active worldwide scanning activities. Other tools such as BinaryEdge, ZoomEye, Fofa, and others are considered for future work.

3.2 Scanner Configuration

Scanner tools utilize specific configurations to be able to query their databases. The configuration plays an important role in the outcome of the scanning process and could lead to the success or failure to meet its objectives. Considering the objective of this scanning process is to fingerprint maritime components worldwide, the configuration should include maritime-specific elements that lead to the desired outcome. Therefore, I propose the utilization of NMEA messages to fingerprint maritime components. The NMEA messages are employed within the queries leading to the identification of possible components. I propose the utilization of the static information in NMEA messages in the fingerprinting process (refer to section 2). Other configurations such as IP range, and ports are not considered relevant in this scanning process.

There are more than a hundred standard types of NMEA messages (i.e. +100 MessageIDs) that can be emitted by more than a hundred types of devices (i.e. +100 TalkerIDs). There is some commonality in the types of messages emitted by certain devices, and some messages are not expected to be emitted by a specific set of devices. However, there are no guidelines that can provide this information. Therefore, I have followed a comprehensive approach for scanning all TalkerID and MessageID pairs in an attempt to scan for all possible NMEA messages. This results in +10000 search queries required to cover all possibilities. A limitation of this approach has been observed during implementation. Some scanner tools limit the number of queries for each user under certain subscription plans. For instance, BinaryEdge and Censys allow for only 250 queries a month for a free subscription. Therefore, for such scanners, I have followed a rather limited yet focused approach for bypassing this issue. My alternative approach is only to query the most common TalkerID and MessageID pairs. Still, I was able to implement the comprehensive approach using Shodan. Additionally, Raymond [26] has compiled comprehensive documentation of NMEA messages which I have relied upon in this paper. Raymond listed a group of uncommon NMEA messages as well as vendor-specific messages. I have included such messages in my scanning scope aiming to achieve comprehensive coverage of NMEA messages. Nevertheless, the standard refers to other vendor-specific messages with a structure that is hard to predict such as starting with the letter "P" or starting with the letter "U" followed by a group of numbers. This means that my coverage of NMEA messages, although comprehensive, it is yet not complete.

The output of both approaches is a group of queries that are used in the next step. To this end, I have developed a group of scripts that can generate all these queries and make them ready to be sent to the Shodan and Censys APIs. The queries are configured to look for banners that include NMEA messages. Table 1 present examples of such queries.

3.3 Scanner Initiation

For this step, I have developed a group of scripts to run all the generated queries against the Shodan and Censys APIs and record the results for analysis. I highlighted the issue of passive scanning with regards to the freshness of results in

Table 1. Examples of query strings

TalkerID	MessageID	Description	Query string
GP	RMC	The static data of a Recommended Minimum Navigation Information (RMC) message emitted by a GPS Device (GP)	”$GPRMC,”
	GGA	The static data of a Global Positioning System Fix Data (GGA) message emitted by a GPS Device (GP)	”$GPGGA,”
GL	HDT	The static data of a Heading - True (HDT) message emitted by a GLONASS Device (GL)	”$GLHDT,”
PGRMZ		A vendor-specific message emitted by Garmin devices containing altitude information.	”$PGRMZ,”

my earlier work [1]. Some queries might return hosts that have been recorded emitting NMEA messages at the time the active scanning was conducted. However, this might not always reflect the correct status of that host. Nevertheless, it has been highlighted by Bennett et al [3] that both Shodan and Censys can reflect updates within 24 hours. Additionally, the IP addresses of the devices emitting NMEA might change overtime, therefore, repeating the scanning process periodically is needed to maintain the most accurate and up-to-date results.

I repeated the search process several times against the Shodan API following the comprehensive approach discussed in section 3.2. However, I followed the focused approach against Censys without repetition due to the limited subscription plan.

3.4 Collection

The query results are stored in files with different formats corresponding to the different scanning tools. I collected records with information including:

- The number of hosts observed to emit each NMEA message. This would shed a light on the most common messages.
- For each observed occurrence of NMEA message by a host, record the host IP, port number or service, country, and banner data containing the message. This information can be utilized for vulnerability analysis and the identification of maritime components. The port numbers as well as the banner data are expected to provide information regarding the device or software that is used for this service. Such information is valuable to attackers during the reconnaissance stage of cyber attacks.
- Record the results of all queries for validation.

3.5 Validation

The validation at this step refers to ensuring that the scanning results are correct and are useful to achieve the scanning objectives. Otherwise, the process is re-initiated with different scanner tools, configuration, or collection approaches. For this use case, it is necessary to validate that the identified hosts are actually emitting NMEA messages.

It is crucial to understand the different employed scanners as each scanner employs a unique query functionality that determines the quality of returned results concerning the scanning objective. For instance, Shodan does not have an "exact match" feature for queries. Instead, queries return results that approximately contain the query string. This has lead to getting false positive matches. The reason behind this is that the string of certain NMEA messages may appear in banners grabbed by hosts but that banner is not relevant to an NMEA service. An observed example of this issue is the query string "$TRACK," which is employed to scan for NMEA message "ACK" (Alarm Acknowledgement) emitted by Talker ID "TR" (TRANSIT Navigation System). Shodan removes the special characters from the query string and the remaining phrase "TRACK" appears in the banner data of many hosts but not as NMEA messages. Therefore, a validation process is required to ensure that only hosts emitting NMEA messages are identified and their data are collected. For this, I have developed scripts that will read all the returned results and only return results that contain correct NMEA messages.

3.6 Analysis

During this step, I analyzed the collected search results discussed in section 3.4. The analysis is different for each scanner tool as each one returns different results with different information. I will highlight the analysis process for the search results obtained from Shodan since it returned the largest amount of results. After removing the duplicate results, I have observed 4992 unique NMEA sessions emitted by 4942 hosts. The session information includes the host IP, port number, country code, banner data as well as a summary of NMEA messages in the banner data. The latter led me to the identification of additional NMEA messages that were outside the scope of the search (refer to section 3.2) but appeared to accompany the messages within the scope. The analysis process included four activities, namely, device identification, general statistics about the NMEA service, maritime component identification, and comparison between the different scanner tools.

Device Identification The identification of IoT devices and their operating system (OS) is among the challenges highlighted in my earlier work [1]. Banner data, port numbers, certificates, and other information have been employed in the literature to identify device types and OSs. This information is afterward employed in the identification and analysis of the vulnerability of such devices and OSs.

I have attempted to identify devices following several approaches. First, a generic classification is possible using the NMEA format. The type of device from which the NMEA message is coming is encoded in the TalkerID. Although the type of IoT device that might be forwarding the messages cannot be identified through this approach, nevertheless, it can shed a light on the type of devices connected to the host. Such information is useful for attackers at the reconnaissance stage. Table 2 depicts the number of detected hosts for each NMEA talker.

The table reflects that the majority of devices are receivers of the major positing systems, namely, Global Positioning System (GPS), GLONASS, and a combination of many systems. Moreover, the quantity of vendor-specific NMEA talkers is observed. Therefore, the second approach relied on vendor-specific messages.

Table 2. Distribution of type of NMEA talkers across the detected hosts

Talkers (Description)	Host Count (%)	Talkers (Description)	Host Count (%)
GP: GPS receiver	4897 (99%)	GB: BeiDou receiver (China)	4 (0%)
Vendor-Specific	1939 (39%)	CC: Computer - Programmed Calculator	2 (0%)
GN: Combination of multiple satellite systems	1595 (32%)	II: Integrated Instrumentation	2 (0%)
GL: GLONASS receiver	1558 (32%)	DF: Direction Finder	1 (0%)
BD: BeiDou receiver (China)	115 (2%)	VW: Velocity Sensor, Speed Log, Water, Mechanical	1 (0%)
GA: Galileo receivere	74 (2%)	SD: Depth Sounder	1 (0%)
AB: Independent AIS Base Station	13 (0%)	YD: Transducer - Displacement, Angular or Linear	1 (0%)
WI: Weather Instruments	10 (0%)	PQ: Quectel Quirk	1 (0%)

Relying on several online resources, I was able to identify some device types known to emit the most common vendor-specific messages based on their TalkerID code. Additionally, I used the National Vulnerability Database (NVD) published by NIST [25] to find possible Common Vulnerabilities and Exposures (CVE) by using the identified device information. I also recorded the CVE's risk ratings that are encoded using the Common Vulnerability Scoring Scheme (CVSS). Table 3 show the identified device types, the number of hosts that emits them, and possible CVEs associated with these devices.

NMEA service In this analysis, I focused on the most observed messages, ports, and countries to stand on the status of NMEA service worldwide. This information is helpful to the cybersecurity community to manage risks related to NMEA.

Regarding message types, I have observed 4 types of AIS messages; ABVDM, AITXT, AIVDO and AIVDM, 41 NMEA messages that are specified in the NMEA-0183 standard [2], 34 messages following the standard specifications for vendor-specific messages, and 29 messages that have no specified description in the standard, however, they have a format similar to NMEA. Table 4 reflects the most observed messages, all of which are standard NMEA messages, brief description, and the number of hosts that emit them. Note that 83,25 % of the observed hosts emit at least two different NMEA messages together. Each

Table 3. The identified device types and some of their possible vulnerabilities

Messages	Description	Host Count (%)	Possible CVEs (CVSS)
Most common: PMTKAGC, PMTKGALM, PMTKGEPH ,PMTKTSX1	MediaTek MTK chipsets	1662 (97,6%)	CVE-2020-13841 (9.8) CVE-2020-13842 (7.8)
PSTT	Saab Systems position receiver	35 (0,9%)	None
PCPTI	Cradlepoint Router	28 (0,7%)	
PLEIR	LEICA GPS receiver	21 (0,5%)	
PTNL	Trimble GNSS Receiver	3 (0,1%)	CVE-2012-5053 (4.2)
PQXFI	Qualcomm chipset	1 (0,0%)	CVE-2021-1965 (9.8) CVE-2021-1955 (7.5)

message provides different valuable information for the Cyber-Physical tracking process. Among the messages in the table, GGA and RMC messages together provide the most amount of information including time, position, speed, heading, and others. Therefore, they are great candidates for fingerprinting maritime components.

Table 4. Top 10 observed NEMA messages emitted through the Internet

Message	Description	Host Count (%)
GGA	GPS Fix Data including position and time information	4815 (96%)
RMC	Recommended Minimum Navigation Information including position, time, speed, and heading.	4145 (83%)
VTG	Track made good and Ground speed	4019 (81%)
GSA	GPS Dilution of precision (DOP) and active satellites	3142 (63%)
GSV	Satellites in view	3077 (62%)
GLL	Geographic Position - Latitude/ Longitude	63 (1%)
ZDA	Time & Date	34 (1%)
GNS	Fix data	16 (0%)
DBT	Depth below transducer	14 (0%)
GST	GPS Pseudorange Noise Statistics	13 (0%)

Regarding used ports, I discovered 92 ports used for transmitting NMEA messages, the majority of which are transmitted through two TCP ports, specifically, port 7000 (50%) and port 50100 (45%). This indicates that scanning these ports alone would cover 95% of the entire NMEA presence on the Internet. Furthermore, I have observed that approximately 10.5% of hosts emitting NMEA messages over a certain port have more than one other port open ranging between 2 to 100 ports including ones used for other services such as HTTP and

Message Queuing Telemetry Transport (MQTT). These other services might as well have their own vulnerabilities. However, my analysis didn't pursue this issue any further.

Regarding countries, in total 66 countries have hosts emitting NMEA messages. The majority of hosts have IP addresses registered in Brazil (78%), Argentina (3,3%), Spain (3,2%), Japan (2,9%), Morocco (2%), and United States (1,7%). Although the high number of NMEA messages in Brazil is unexpected, I have not investigated the reason behind any further in this paper.

Maritime components identification As mentioned in section 1, one of the objectives of this paper is to investigate the ability of attackers to fingerprint maritime components and identify their cyber-related information as well as physical information during the reconnaissance stage to aid during further stages of cyber attacks. In this section, I present my method and results for the identification of maritime components with observed presence on the internet, identify their cyber-related information (IP, ports, data, and vulnerabilities), and track their physical location to obtain combined cyber-physical records of the components. My approach relies on the following assumption, a host is considered a maritime component under two conditions, its communicated coordinates are located at a sea area or it is emitting an NMEA message with a talker that is an AIS base station.

Based on that, I developed an algorithm that will parse the NMEA banner data for each host, and obtain valid coordinates information (latitude and longitude) from either RMC or GGA messages. Then I utilized an algorithm provided by Karin [12] to check if these coordinates belong to a sea or land area. Moreover, if an AIS base station Talker ID is observed, a component is labeled as a possible maritime component. The results of the algorithm are depicted in Table 5. I have detected 331 possible maritime components, obtained their cyber as well as physical information. To evaluate my algorithm. I have manually and randomly verified some of the obtained results. I have randomly chosen 10 detected land positions, 10 sea positions and checked; using Google Maps, if they are accurately labeled. The results suggests that my algorithm returns valid results. Further development and evaluation are expected for future work by utilizing vessel tracking services to correlate the cyber and physical properties.

Table 5. Results of the maritime component fingerprinting algorithm

Maritime Component?	Rational	Count
No	No Evidence	159
No	Land position	4502
Yes	Sea position	325
Yes	AIS Base Station	6

Comparison between scanners In this section, a comparison is presented for the two most referenced scanners, namely, Shodan and Censys concerning the scanning process in this paper. Table 6 depicts a summary of this comparison. Shodan provided the best possible results for analysis due to a sufficient subscription plan. Therefore, this comparison doesn't reflect the actual utility of each scanner. However, it justifies the focus of the analysis on the records collected from Shodan.

Table 6. Comparison between NMEA queries between Shodan and Censys

	Scanner	**Initiated Queries**	**Messages Detected**	**Collected Records**
	Censys	182	52	9582
	Shodan	12206	53	22726
	Shodan more	**Censys more**	**Same results**	**Both 0**
# of Messages	34	18	5	99

4 Discussion and Limitations

Discussing the risks of the Cyber-Physical tracking process can be conducted by considering the risks of the associated *ATT&CK* techniques. *ATT&CK* [29] is a common knowledge repository for observed cyber adversarial behaviors. The presented Cyber-Physical tracking process in this paper emulates an adversarial behavior that includes several techniques indicated in the *ATT&CK* framework. The relevant techniques to this paper are i) Gather Victim Host Information (T1592) [20], ii) Search Open Technical Databases (T1596) [22], and iii) Search Open Websites/Domains (T1593) [23]. In my approach, I have fine-tuned the scanning process by searching open websites and domains as resources for identifying information such as vendor-related information. Also, I have searched technical databases such as Shodan, Censys, and NVD. Moreover, I have gathered the victim host information such as IP address, ports, possible device type as well as possible vulnerabilities. The IP addresses and ports can later be utilized for subsequent adversarial techniques to gain initial access to the victims' networks or impact the operations of the emitting devices. Initial access might later be achieved through External Remote Services (T1133) [18]. Considering that the NMEA messages were detected from the internet, this indicates that each emitting device has at least one external-facing remote service that is remotely accessible. Moreover, as mentioned in Section 3.6, 10.5% of the detected NMEA-emitting hosts have between 2 to 100 remote services open. If any of these services has a vulnerability that can be remotely exploited, it may lead to enabling the attacker to gain an initial foothold to the connected network. Additionally, attackers may attempt to inflect impact through remote Network

Denial of Service (T1498) [19] in the case that the NMEA talker is susceptible to such vulnerability.

The *ATT&CK* framework refers to the difficulty of mitigating the techniques T1592, T1596, and T1593 as they are performed outside the scope of the defensive capabilities of organizations. However, efforts to minimize the availability and sensitivity of data to external parties are suggested. The *ATT&CK* framework mentions the very high occurrence and associated false positive rate of such activities. This is reflected in my work as the ability to scan is always possible even without a proper subscription. Additionally, a certain false positive rate is expected due to the passive scanning approach. The obtained results might not reflect the actual status of hosts. However, the obtained NMEA messages from the banner data include several fields containing the Coordinated Universal Time (UTC) the information was captured which can reflect the freshness of the scanning result. Additionally, it has been reported that both Shodan and Censys reflect updates within 24 hours [3].

Similar adversarial techniques have been observed by the cybercriminal group "Sandworm Team" [21] during the development of the NotPetya attack. The Sandworm team searches open websites and databases for information to craft credible spearfishing emails [4]. This incident highlights the utility of such available resources to attackers and the necessity to investigate such threats in different domains including maritime.

The maritime sector is witnessing a digital transformation era leading to expected drastic changes in technology, business models and operations [10]. A new operational model for future maritime components has been communicated by members of the classification society in maritime, specifically, the Norwegian organization DNV. The operational mode is called auto-remote; autonomous as possible and remotely controlled when needed [7]. Tam and Jones [30] have discussed the unique cyber-physical opportunities in specific geological locations when considering futuristic unmanned ships. The authors indicated the utility of such opportunities to pirates adopting cyber attack techniques. Therefore, the associated risks of the demonstrated approach in this paper are increased when considering the auto-remote operational mode.

The field of IoT vulnerability scanning is recent and growing [1]. Also, the field of Cyber-physical tracking is scarce as very limited artworks have discussed it. However, I argue that immediate actions are needed for demonstrating the feasibility and possible impacts of such activities. This is important to support the ongoing efforts for improving cybersecurity in maritime. Therefore, I acknowledge the following limitations in the proposed approach and discuss the rationals for dealing with them:

- The choice for utilizing Shodan is only to present a proof of concept for the proposed approach. Other scanner tools might provide different results. For instance, it has been communicated by Li et al [13] that ZoomEye scans over 1.2 billion devices compared to only 0.4 by Shodan. Therefore, evaluating the proposed approach using ZoomEye and other scanner tools is considered for future work.

- The scope of the utilized NMEA messages in the fingerprinting process is limited by the discussed messages in the NMEA-0183 standard [2] as well as the comprehensive documentation of NMEA protocol by Raymond [26]. Other messages that are not documented or that have not appeared in the search might exist but are still undetected. However, the results suggest that only a few of the NMEA messages types constitute the majority of the detected messages which might render the impact of any missing messages insignificant.

5 Conclusion

Cyber security in the maritime domain is a growing area of interest due to the undergoing digital transformation. The maritime infrastructure is consuming additional digital components including IoT and Industrial IoT [15, 5]. The future of the maritime domain includes new modes of operation (e.g. auto-remote) that require increased connectivity and reduces human presence around maritime components. Risk management activities in maritime have been proposed by the International Maritime Organization (IMO) and documented in Resolution MSC. 428(98) [6]. Such activities include analyzing the threat landscape and continuous improvement of defenses.

This paper supports the efforts in this direction as it demonstrates an offensive capability that can be conducted by attackers to gain tactical advantages by combining cyber and physical information regarding maritime components and utilize them during the development of more directed cyber-physical attacks. I have presented a new approach for scanning and identifying cyber-related information for NMEA emitting devices. The scanning process yielded in identifying 4942 hosts emitting NMEA messages the majority of which using ports 7000 and 50100. I have also identified several device types and expected vulnerabilities. Such information aims to capture the status of NMEA service worldwide to attract attention toward improved cybersecurity.

Additionally, I have proposed a new approach for detecting maritime components that are connected to the internet. The algorithm utilized information collected from IoT scanners of hosts emitting NMEA messages some of which include position information. The algorithm detected 331 maritime components that are connected to the internet. It identifies their location, speed, and other physical properties in addition to their IP addresses, ports, and other cyber properties. Such components could be susceptible to cyber-physical attacks. In summary, I argue that the Cyber-Physical tracking process constitutes a threat against the detected maritime components and I urge the maritime community to consider the outcome of this work.

References

1. Amro, A.: Iot vulnerability scanning: A state of the art. Computer Security pp. 84–99 (2020)

2. Association, N.M.E., et al.: Nmea0183 standard. =https://www.nmea.org/content/STANDARDS/NMEA$_0$183$_standard(2002)$
3. Bennett, C., Abdou, A., van Oorschot, P.C.: Empirical scanning analysis of censys and shodan
4. Brady., S.W.: United States vs. Yuriy Sergeyevich Andrienko et al (2020, October 15), https://www.justice.gov/opa/press-release/file/1328521/download
5. Chubb, N.: Maritime Applications for IoT ((accessed September 8, 2021)), https://thetius.com/maritime-applications-for-iot/
6. Committee, T.M.S.: International maritime organization (imo) (2017) guidelines on maritime cyber risk management. http://bit.ly/MSC428-98
7. DNV GL: Dnvgl-cg-0264: Autonomous and remotely operated ships (2018)
8. Durumeric, Z., Adrian, D., Mirian, A., Bailey, M., Halderman, J.A.: A search engine backed by internet-wide scanning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. pp. 542–553 (2015)
9. Fiorini, M.: Maritime awareness through data sharing in vts systems. In: 2012 12th International Conference on ITS Telecommunications. pp. 402–407. IEEE (2012)
10. Fruth, M., Teuteberg, F.: Digitization in maritime logistics—what is there and what is missing? Cogent Business & Management **4**(1), 1411066 (2017)
11. Haynes, D.: Iran's secret cyber files on how cargo ships and petrol stations could be attacked (Jul 2021), https://news.sky.com/story/irans-secret-cyber-files-on-how-cargo-ships-and-petrol-stations-could-be-attacked-12364871
12. Karin, T.: Global land mask. https://github.com/toddkarin/global-land-mask (October 2020)
13. Li, R., Shen, M., Yu, H., Li, C., Duan, P., Zhu, L.: A survey on cyberspace search engines. In: China Cyber Security Annual Conference. pp. 206–214. Springer, Singapore (2020)
14. Luft, L.A., Anderson, L., Cassidy, F.: Nmea 2000 a digital interface for the 21st century. In: Proceedings of the 2002 National Technical Meeting of The Institute of Navigation. pp. 796–807 (2002)
15. Maritime, T.: The Internet of Things Makes Waves on a Global Maritime Network ((accessed September 8, 2021)), https://telenormaritime.com/digital-shipping/internet-of-things-iot/
16. Matherly, J.: Complete guide to shodan. Shodan, LLC (2016-02-25) **1** (2015)
17. Miessler, D., Smith, C.: Owasp internet of things project. OWASP Internet of Things Project-OWASP (2018)
18. MITRE: External Remote Services (T1133) (2021 (accessed November 2, 2021)), https://attack.mitre.org/techniques/T1133
19. MITRE: Network Denial of Service (T1498) (2021 (accessed November 2, 2021)), https://attack.mitre.org/techniques/T1498/
20. MITRE: Gather Victim Host Information (T1592) (2021 (accessed September 8, 2021)), https://attack.mitre.org/techniques/T1592
21. MITRE: Sandworm Team (2021 (accessed September 8, 2021)), https://attack.mitre.org/groups/G0034/
22. MITRE: Search Open Technical Databases (T1596) (2021 (accessed September 8, 2021)), https://attack.mitre.org/techniques/T1596
23. MITRE: Search Open Websites/Domains (T1593) (2021 (accessed September 8, 2021)), https://attack.mitre.org/techniques/T1593
24. Munro, K.: Tracking hacking ships with shodan ais (Jan 2018), https://www.pentestpartners.com/security-blog/tracking-hacking-ships-with-shodan-ais/

25. NVD, N.: National vulnerability database (2011)
26. Raymond, E.S.: https://gpsd.gitlab.io/gpsd/NMEA.html
27. Sivkov, Y.: Transformation of nmea ship network from sensor-based to information-based model. In: 2018 20th International Symposium on Electrical Apparatus and Technologies (SIELA). pp. 1–4. IEEE (2018)
28. statista.com: Internet of things (iot) connected devices installed base worldwide from 2015 to 2025. https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/
29. Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G., Thomas, C.B.: Mitre att&ck: Design and philosophy. Technical report (2018)
30. Tam, K., Jones, K.: Cyber-risk assessment for autonomous ships. In: 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). pp. 1–8. IEEE (2018)
31. Tran, K., Keene, S., Fretheim, E., Tsikerdekis, M.: Marine network protocols and security risks. Journal of Cybersecurity and Privacy **1**(2), 239–251 (2021)
32. Tundis, A., Mazurczyk, W., Mühlhäuser, M.: A review of network vulnerabilities scanning tools: types, capabilities and functioning. In: Proceedings of the 13th International Conference on Availability, Reliability and Security. pp. 1–10 (2018)

Generation of Non-Deterministic Synthetic Face Datasets Guided by Identity Priors

Marcel Grimmer[1], Haoyu Zhang[1], Raghavendra Ramachandra[1], Kiran Raja[1], and Christoph Busch[1,2]

[1] NBL - Norwegian Biometrics Laboratory, NTNU, Norway
[2] da/sec - Biometrics and Internet Securiy Research Group, HDA, Germany

Abstract. Enabling highly secure applications (such as border crossing) with face recognition requires extensive biometric performance tests through large scale data. However, using real face images raises concerns about privacy as the laws do not allow the images to be used for other purposes than originally intended. Using representative and subsets of face data can also lead to unwanted demographic biases and cause an imbalance in datasets. One possible solution to overcome these issues is to replace real face images with synthetically generated samples. While generating synthetic images has benefited from recent advancements in computer vision, generating multiple samples of the same synthetic identity resembling real-world variations is still unaddressed, i.e., mated samples. This work proposes a non-deterministic method for generating mated face images by exploiting the well-structured latent space of StyleGAN. Mated samples are generated by manipulating latent vectors, and more precisely, we exploit Principal Component Analysis (PCA) to define semantically meaningful directions in the latent space and control the similarity between the original and the mated samples using a pre-trained face recognition system. We create a new dataset of synthetic face images (SymFace) consisting of 77,034 samples including 25,919 synthetic IDs. Through our analysis using well-established face image quality metrics, we demonstrate the differences in the biometric quality of synthetic samples mimicking characteristics of real biometric data. The analysis and results thereof indicate the use of synthetic samples created using the proposed approach as a viable alternative to replacing real biometric data.

Keywords: Biometrics, Face recognition, Synthetic Face Image Generation, Deep learning, StyleGAN

1 Introduction

The popularity of biometric recognition has increased steadily along with the development of more accurate and convenient recognition technologies. According to ISO/IEC 2382-37:2017 [17], biometrics refers to the automated recognition of individuals based on their biological and behavioural characteristics. In particular, the human face has proven to be sufficiently unique and an easy-to-capture biometric characteristic, leading to a wide range of real-world applications, including border control, passport issuance, and civilian ID management. Driven by the promising performance of current face recognition systems, the Smart Borders program has been initiated

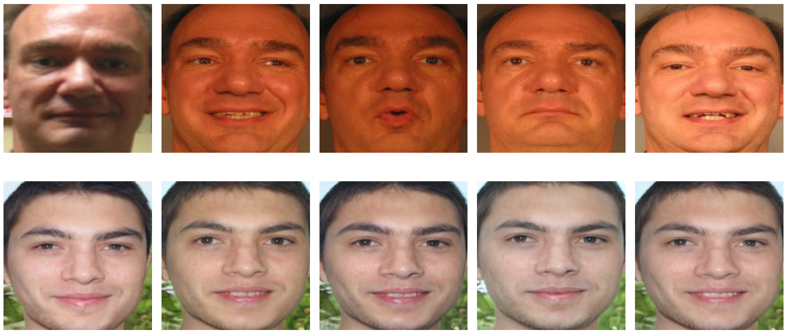


Fig. 1: Comparison of the intra-identity variation between FRGC v2.0 (top) and SymFace (bottom).

within the European Union to establish the Entry-Exit System (EES) [4], an automated IT system for registering travellers from third-countries, replacing the current system of manual stamping of passports. This system aims to help bona fide third-country nationals travel more easily while also identifying more efficiently over-stayers and cases of document and identity fraud. To perform automatically, EES will register the person's name, type of the travel document and biometric data (face images and/or fingerprints).

A requirement for deploying biometric recognition at the European borders is complying with the high standards defined in the best practices for automated border control of the European Border and Coast Guard Agency (Frontex) [6]. The compliance with these guidelines must be validated by conducting large-scale biometric performance tests which require large datasets. As the collection of real face images is expensive, time-consuming, and privacy-concerning, generating synthetic face images has become an attractive and viable alternative. Driven by the advancements in technology, approaches like StyleGAN and StyleGAN2 [20][21] have shown promises to create large scale face datasets with unique identities.

While the synthetic image generation approaches are well used in various applications, the applicability of those images in biometrics is limited. Specifically, the biometric data used for training algorithms and performance testing need to mimic the real data with variations in pose, varying expressions, occlusions and illumination changes reflecting realistic conditions for any particular identity. In essence, each synthetic identity should accompany a set of variations that can compose what is referred to as mated samples for obtaining comparison scores. Specifically, the synthetic data should represent intra-class variations similar to bona fide data while preserving the identity information. The mated samples essentially are required to generate the genuine score distribution to gauge the biometric performance such as False Match Rate (FMR) and False Non-Match Rate (FNMR). However, despite the recent advancements of synthetic image generation [20][21], it continues to be a technical issue to create synthetic datasets with mated samples that are representative and comparable to real face images captured at border control scenarios (e.g. frontal head poses without face occlusions).

1.1 Our contributions

This work tackles the above-described challenge by introducing a new technique for generating synthetic mated samples. More precisely, a pre-trained StyleGAN generator [20] is utilised to generate synthetic face images of distinct synthetic individuals ("base images"). Each base image is represented by a latent vector $w_{1\times512}$, acting as a compressed version of the original image and reflecting the internal data representation learned by StyleGAN. Motivated by the idea of editing facial attributes by shifting the corresponding latent vector in a specific direction in the latent space [25], we propose to generate mated faces in a non-deterministic manner. We assert that such an approach for attribute editing leads to a better approximation of the natural intra-identity variation of bona fide mated samples, as can be compared in Figure 1.

As the components of the latent vector space can represent various possible semantics, the principal components can be interpreted as semantically meaningful directions in the latent space of StyleGAN. Concretely, extracting the Principal Components [22] from a latent vector space of $50,000$ to 512 leads to obtaining semantically meaningful values. Inspired by such an argument, we create the mated samples by shifting the latent vectors into the directions given by the most relevant eigenvectors (i.e. the principal components). However, as the latent vectors are moved farther from their original locations, the risk of losing the identity information increases, we, therefore, employ a pre-trained face recognition system (FR) [5] to obtain the distance between the original and edited image dynamically to ensure the preservation of identity information from mated samples for the original identity used for editing. We refer to non-deterministic face editing as changing multiple semantics in an unsupervised manner, as opposed to controlled face editing, where specific facial attributes are chosen to be edited.

With such a rationale of our proposed approach, we create a new dataset of face images with synthetic identities and mated samples for each identity in this work which we refer to as Synthetic Mated Face Dataset (SymFace Dataset). The dataset consists of $77,034$ samples with an average number of three mated samples per synthetic identity. To better approximate a semi-controlled capturing environment, images with extreme characteristics are sorted out, taking into account illumination conditions, head poses rotation and inter-eye distance. Also, the study concentrates on adult face images due to the limited training data available from young children and seniors. We refer to Figure 3 to get an impression of typical images filtered out by our filtering pipeline.

We further evaluate the quality of our proposed synthetic dataset by comparing its properties to real face images taken from FRGC v2.0 [23]. Among other approaches for conducting such an analysis, we translate the biometric quality of each image to a quality score between [0, 100] using Face Quality Assessment Algorithms (FQAAs) [19]. In this context, a high-quality score indicates that the corresponding biometric sample is well suited for biometric recognition. On the opposite, low-quality scores deteriorate the recognition accuracy due to the low quality of the input image. This understanding of biometric quality corresponds to the terminology specified by ISO/IEC 29794-1 [16], defining the utility of a biometric sample as the prediction of the biometric recognition performance. In this work, two FQAAs are used for estimating and comparing the biometric quality: FaceQnet v1 [11] and SER-FIQ [26]. At this point, the reader is referred to Section 2 to obtain a more detailed description of these methods.

In the rest of the paper, Section 2 summarises the conceptual ideas of generating synthetic face images and mated samples. Next, Section 3 provides a detailed description of the proposed PCA-FR-Guided sampling approach. Section 4 details the newly created SymFace Dataset, and finally, Section 5 gives an overview of the experimental results, followed by a conclusion about the key findings in Section 6.

2 Related Works

2.1 Synthetic Image Generation

In 2019, Karras et al. [20] presented a style-based generator architecture for generative adversarial networks (StyleGAN), capable of generating synthetic images with high resolutions (1024x1024) and realistic appearances. In addition to their proposed GAN architecture, the authors web crawled high-quality human face images from a social media platform (Flickr) to create a new dataset (FFHQ), covering a wide variation of soft biometrics.

Despite the recent success of deep generative networks, most generators are still operating as black-boxes and lack a deeper understanding of the latent space. To address these weaknesses and improve the disentanglement properties of the latent space, StyleGAN maps initially drawn latent vectors to an intermediate latent space, which turns out to encode facial features in a more disentangled manner. Further, Adaptive Instance Normalization (AdaIN) [13] enables to fuse the styles of different faces on multiple feature levels. Furthermore, stochastic variation is achieved by adding Gaussian noise to the feature maps after each convolution operation to vary fine-grained details. Recently, StyleGAN2 has been published by the same authors [21], improving the architectural design and fixing the characteristic artefacts occurring in the synthetic images generated by StyleGAN.

In StyleGAN and StyleGAN2, synthetic images are generated by randomly sampling from a known distribution (latent space). If these latent vectors are drawn from tail regions of the distribution, the quality of the generated face images deteriorates while the diversity of facial attributes increases. To balance this trade-off, a truncation factor can be used to stabilise the sampling: the truncated latent code w' is calculated as $w' = \bar{w} + \psi(w - \bar{w})$ where $\bar{w}$ indicates the latent spaces' center of mass and ψ denotes the truncation factor. Following the empirical analysis of Zhang et al. [9], we choose a truncation factor of $\psi = 0.75$. In [9], the authors have shown that the biometric performance of synthetic samples generated with StyleGAN and StyleGAN2 are similar and comparable to bona fide images from FRGC v2.0 [23]. Hence, this work uses StyleGAN for generating synthetic base images to enable the implementation of PCA-FR-Guided sampling to operate within the framework of InterFaceGAN [25].

2.2 Mated Sample Generation

Though it has been shown in [9] that single synthetic face images can achieve comparable performances as bona fide samples for face recognition, mated samples are more commonly required in biometric performance evaluations. Given a synthetic base image, mated samples can be derived by editing facial attributes to simulate the factors of variation present in bona fide samples. With the groundbreaking work of Shen et al. [25], InterFaceGAN was introduced as a framework enabling editing facial attributes of synthetic identities through manipulating latent vectors in the latent space. In this context, the latent space reflects the internal data representation of

StyleGAN and structures various semantics learned from the training dataset. Further, the innovative architecture of StyleGAN significantly reduces the entanglement of the encoded semantics, which provides optimal conditions for controlled modifications on facial attributes.

The main contribution of InterFaceGAN is based on the observation that the latent space can be divided into linear subspaces according to binary semantics, such as "smile" or "no smile". Concretely, linear Support Vector Machines (SVMs) [3] are used to divide the latent space into subspaces for each facial attribute of interest. Once the SVMs are trained, facial attributes are modified by shifting the latent vectors into the perpendicular direction of the previously found boundaries, thereby causing continuous changes. The same principle has been adopted by Colbois et al. [2], who manipulate yaw angle, illumination, and a smile by approximating the bona fide conditions of Multi-PIE [7].

2.3 Limitations in State-of-the-art

Although InterFaceGAN generates visually appealing mated samples, their applicability for general biometric performance tests is still limited and understudied. As shown in Figure 1, mated samples collected in real-world scenarios naturally include several variations varying at the same time, for instance, pose, illumination, expression and a combination of them. In contrast, controlled face editing focuses on changing only a few semantics while leaving others fixed. Therefore, controlled modifications are useful to determine the vulnerability of face recognition systems for targeted semantics while only representing a small subset of the potential diversity in bona fide datasets. Motivated by this observation, we introduce PCA-FR-Guided sampling as a technique for generating non-deterministic mated samples to either replace or complement existing test datasets.

3 PCA-FR-Guided Sampling

This section introduces our new method for generating mated samples, which we refer to as PCA-FR-Guided Sampling. As described in Section 2, semantic modifications can be caused by moving latent vectors in the latent space. However, this approach still leaves two questions unanswered: 1) How to choose semantically meaningful directions? 2) How to choose the distance to preserve identities while maximising the intra-identity variation?

Aiming to find solutions for the aforementioned questions, Figure 2 provides an overview of the PCA-FR-Guided sampling technique. After generating an initial synthetic dataset with StyleGAN with a truncation factor of $\psi = 0.75$ (A), PCA is applied to extract semantically meaningful directions from the corresponding latent vectors (B). The idea is to extract the latent direction with the most variance, leading to effective variation after image generation. Finally, the latent vectors are moved along the principal component axes while adjusting the distance dynamically by measuring the similarity between the original and the shifted mated sample in a step-wise manner (C). Algorithm 1 provides a detailed workflow of the PCA-FR-Guided mated sample generation process proposed in this work.

We specifically employ *stepSize* and the verification *threshold* as controlling parameters to balance the trade-off between intra-class variation and identity-retaining factor for generated mated samples. In other words, increasing the comparison thresh-

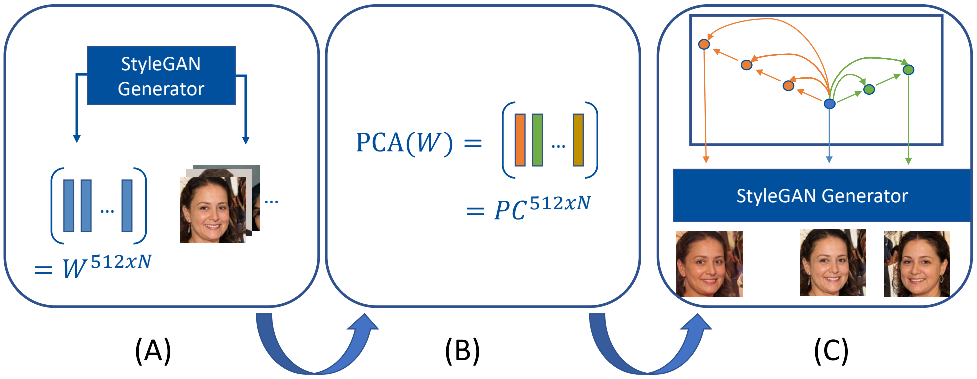


Fig. 2: Overview of the proposed PCA-FR-Guided Sampling with $N = 50,000$ denoting the number of latent vectors concatenated as matrix W to obtain the principal components (PCs). A detailed workflow is given in Algorithm 1.

old decreases the distance between the original latent vector and the shifted latent vector, thus generating more similar faces with fewer factors of variation. On the other hand, decreasing the step size approaches the given threshold with smaller steps, thus yielding mated samples closer to the desired similarity tolerance [3].

4 Synthetic Mated Face (SymFace) Dataset

This section describes the structure of our synthetic mated face dataset (SymFace) and the reference dataset used for the comparison part in Section 5.

Each mated sample is generated based on a synthetic face image randomly generated by StyleGAN. As StyleGAN was trained using images crawled from social media, the diversity of the generated images roughly corresponds to approximate capturing scenarios "in the wild". As described in section 2.1, a truncation factor of $\psi = 0.75$ was chosen to generate 50,000 unique identity images with high resolutions of 1024x1024 pixels and this is referred to as base images.

However, not all images generated from StyleGAN satisfy the minimum criteria needed for biometric applications. For instance, in a border-crossing scenario, factors such as minimum inter-eye distance (IED), illumination metrics, predicted head poses [1], and estimated ages [27] are needed in accordance to ISO/IEC TR 29794-5:2010 [15] and ICAO 9303 [14]. Accounting for this, we discard all such images not meeting the criteria of minimum inter-eye distance (IED), illumination metrics, predicted head poses [1], and estimated ages [27]. The SymFace Dataset thus has a total of 25,919 images which we deem as usable for further analysis in this work, and a sample of such images that are eliminated by our filtering pipeline is illustrated in Figure 3. As it can be observed from Figure 3, despite these images looking visually pleasing, they fail to meet the quality standards with respect to ISO/IEC TR 29794-5:2010 [15] and ICAO 9303 [14].

[3] We have chosen $stepSize = 0.2$ and $threshold = 0.8$ empirically, considering the quality of the mated samples and the algorithm's efficiency. However, other values can also be used on application scenarios.

```
input: $latentVectors$, $N$, $stepSize$, $threshold$, $Generator$
$components = PCA(latentVectors)$;
for $w$ in $latentVectors$ do
    $img = Generator(v)$;
    for $c$ in $components$ do
        $i = 1$;
        do
            $w\_moved = shift\_in\_lspace(w, c, stepSize \cdot i)$;
            $mated\_img = Generator(w\_moved)$;
            $recognised = ArcFace(img, mated\_img, threshold)$;
            $i = i + 1$;
            if $recognised$ then
                $save(mated\_img)$;
        while $recognised$;
    end
end
```

Algorithm 1: PCA-FR-Guided sampling algorithm for generating mated-samples.

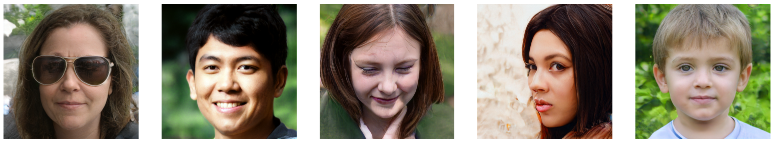


Fig. 3: Low quality images filtered out by our filtering pipeline - from left to right: IED, illumination, pitch angle, yaw angle, age.

Finally, the filtered base images are used as a basis for generating two mated samples for each synthetic identity by using our proposed PCA-FR-Guided sampling technique. Though we selected the first and second principal components, our experiments indicate that each of the 512 principal components can be used to obtain semantically meaningful mated samples. In addition, we apply InterFaceGAN to create three additional datasets, each of which includes mated samples with single semantics edited (yaw angle, illumination quality, and smile).

	SymFace	FRGC v2.0	Illumination Quality	Smile	Yaw
# Base Images	50,000	/	50,000	50,000	50,000
- Filtering	25,919	/	25,919	25,919	25,919
+ Mated Samples	77,757	24,025	77,757	77,757	77,757
- Filtering	77,034	17,919	74,183	74,574	60,504

Table 1: Dataset sizes in different development stages after applying our filtering pipeline and generating mated samples.

4.1 Reference Biometric Dataset

Further, we employ FRGC v2.0 [23] as a reference dataset, including 24,025 bona fide images captured in constrained conditions that resemble the image quality in a border-crossing scenario. Finally, we analyse biometric use cases of the SymFace dataset by studying the characteristics and comparing the same against the FRGC v2.0 dataset. A concise overview of the above-described datasets is given in Table 1, listing the number of samples counted during different development stages. Moreover, Figure 4 presents example images extracted from all datasets, annotated with quality scores obtained by SER-FIQ and FaceQnet v1.

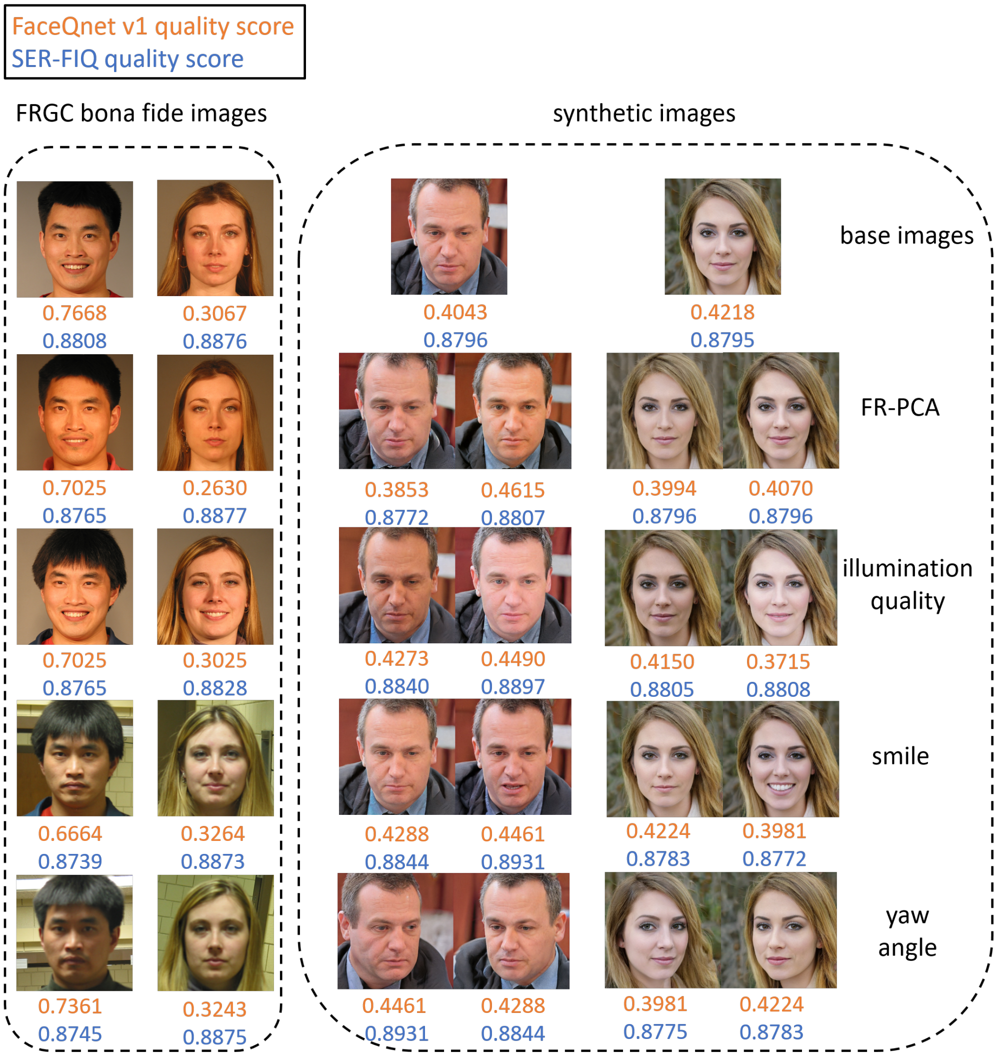


Fig. 4: Examples images of bona fide and synthetic images evaluated in Section 5.

5 Experimental Results

The biometric utility of the synthetic database, especially for mated samples, can be evaluated by measuring the biometric performance or by validating the quality of the samples according to well-established face image quality metrics. We employ both approaches by first evaluating the Face quality assessment algorithms (FQAAs) on the newly created SymFace Dataset and compare it against similar characteristics of the FRGC v2.0 dataset. We then evaluate the mated and non-mated comparison score distributions obtained by applying the pre-trained VGGFace2 [24] face recognition model to verify the biometric utility by analysing the score distribution. We provide a summary of the employed FQAAs for the convenience of the reader.

5.1 Face Image Quality Assessment

Face quality assessment algorithms (FQAAs) are used as indicators of how the quality of a face image contributes to the overall accuracy of a face recognition system. In this work, two representative FQAAs are utilised to evaluate the generated mated samples' biometric quality:

- **FaceQnet v1** is a deep learning-based FQAA proposed by Hernandez-Ortega et al. [11], aiming to predict the general utility of a face image, independent of a specific face recognition system. For the quality score prediction, a pre-trained network of ResNet-50 [10]is fine-tuned as a feature extractor on a small subset of the VGGFace2 dataset [24], including 300 subjects. FaceQnet v1 follows a supervised learning approach, which means that the ground truth quality scores are required for fine-tuning the model. However, finding representative quality scores that accurately reflect general utility criteria is a challenging task. Therefore, the authors propose to determine the utility of an image by comparing it to an ICAO 9303 [14] compliant image, knowing that the sample with unknown image quality can only cause low comparison scores. The performance of FaceQnet v1 has been benchmarked against other FQAAs and proven competitive in the ongoing quality assessment evaluation of the National Institute of Standards and Quality (NIST)[8].
- **SER-FIQ** [26] is an unsupervised technique that is not dependent on previously extracted ground truths for training a prediction model. Compared to FaceQnet v1, which outputs the general utility of a face image, SER-FIQ focuses on predicting the utility for a specific face recognition system. More precisely, the quality scores are based on the variations of face embeddings stemming from random subnetworks of a face recognition model. The authors argue that a high variation between the embeddings of the same sample functions as a robustness indication, which is assumed to be synonymous with image quality. The computational complexity of SER-FIQ increases quadratically with the number of random subnetworks, which leads to a trade-off between the efficiency of the algorithm and the expected accuracy of the quality predictions. In this work, we are following the authors' recommendation, choosing $N = 100$ stochastic embeddings. The comparison of the authors against state-of-the-art FQAA approaches indicates that SER-FIQ significantly outperformed alternative methods.

The distributions of the quality scores predicted with FaceQnet v1 and SER-FIQ are shown in Figure 5. On the left, the well-aligned curves indicate that the average biometric quality across the evaluated datasets is nearly identical. However, looking at the SER-FIQ quality scores reveals a discrepancy between the distributions of the synthetic and bona fide images. We explain this observation with a wider range of yaw angles present in the synthetic datasets, a factor known by the authors of SER-FIQ to decrease the utility estimations [26]. The same behaviour is reflected by the left-shifted purple curve, thereby validating the negative impact of yaw angle variations on the biometric quality. Overall, the analysis of the utility scores does not reveal significant differences between bona fide and synthetic images. Moreover, except for yaw angle manipulations, these differences even vanish when comparing only synthetic datasets. Hence, the biometric quality of mated samples generated with PCA-FR-Guided sampling and InterFaceGAN are similar as both are products of the same generator. Further, the generation of mated samples has not deteriorated the biometric quality, as indicated by the overlapping areas to the base image distributions.

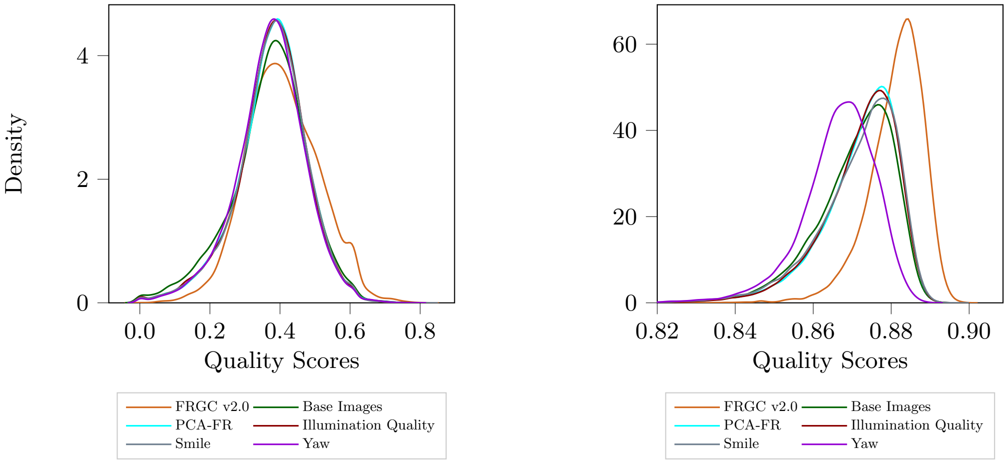


Fig. 5: Quality score distributions of two FQAAs: FaceQnet v1 (left) and SER-FIQ (right).

To further investigate the credibility of the FQAAs, Error-vs-Discard Characteristic (EDC) curves are shown in Figure 6. EDCs are commonly used to compare the performance of multiple FQAAs as suggested by the third version of ISO/IEC WD 29794-1:2021 [18]. For each face comparison, a paired quality score is defined as the minimum of the single quality scores predicted with the FQAAs. Finally, EDC curves are obtained by measuring the FNMRs by increasingly discarding the lowest quality images from the test set. Hence, decreasing EDC curves indicate lower misclassification rates; thus, the underlying FQAA could predict the biometric quality.

In Figure 6, all EDC curves share the same decreasing trend. However, the orange curves (FRGC v2.0) are steeper, indicating that both FQAAs are more accurate in predicting the biometric quality of bona fide images than synthetic samples. One reason for this observation might be rooted in an increased intra-identity variation of the bona fide images, which is still challenging to mimic with synthetic substitutes. This assumption fits with the analysis of the mated comparison scores presented in the following subsection.

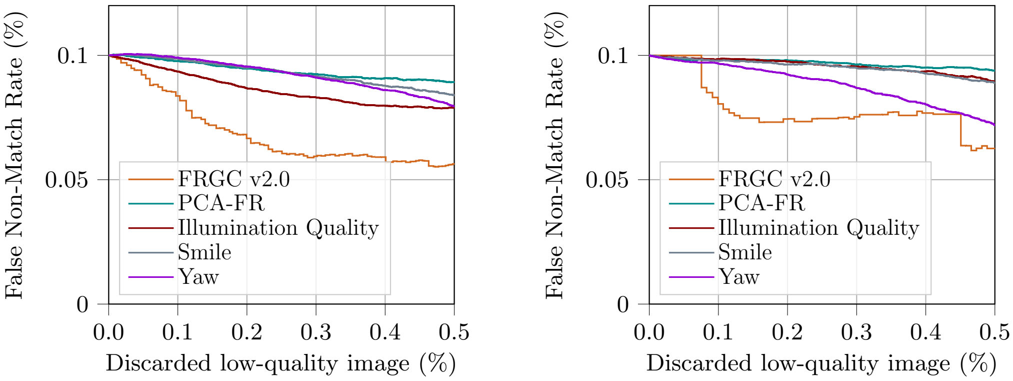


Fig. 6: EDC curves based on paired quality scores derived with FaceQnet v1 (left) and SER-FIQ (right). False non-match rates are computed with ArcFace [5].

5.2 Biometric Performance

After evaluating the biometric quality with dedicated FQAAs, the synthetic face images are further assessed in Figure 7, visualising the comparison score distributions. The non-mated comparison score distributions of all synthetic datasets are clearly below the vertically marked threshold, therefore indicating that the proportion of non-mated lookalikes is not significantly increased compared to non-mated bona fide samples. However, the mated comparison scores reveal more significant differences between the datasets. It is visible that the thick orange curve (FRGC v2.0) is heavier tailed on the left side than all synthetic mated distributions. Again, this observation re-validates the findings of the last subsection, tracing back the differences to a lower similarity of mated samples caused by varying facial attributes. Moreover, the mated comparison scores of the synthetic datasets reveal minor differences: While our proposed PCA-FR-Guided sampling performs similar to the controlled manipulation of smiles, editing yaw angles widens the span of mated comparison scores significantly. Overall, the well-separated distributions of mated and non-mated comparison scores lead us to conclude that either editing single (InterFaceGAN) or multiple (PCA-FR-Guided sampling) semantics can be promising for generating synthetic datasets for biometric performance tests. In addition, a quantitative analysis, measuring the Kullback-Leibler Divergences between the distributions presented in this section, is provided in Table 2 and Table 3 (appendix).

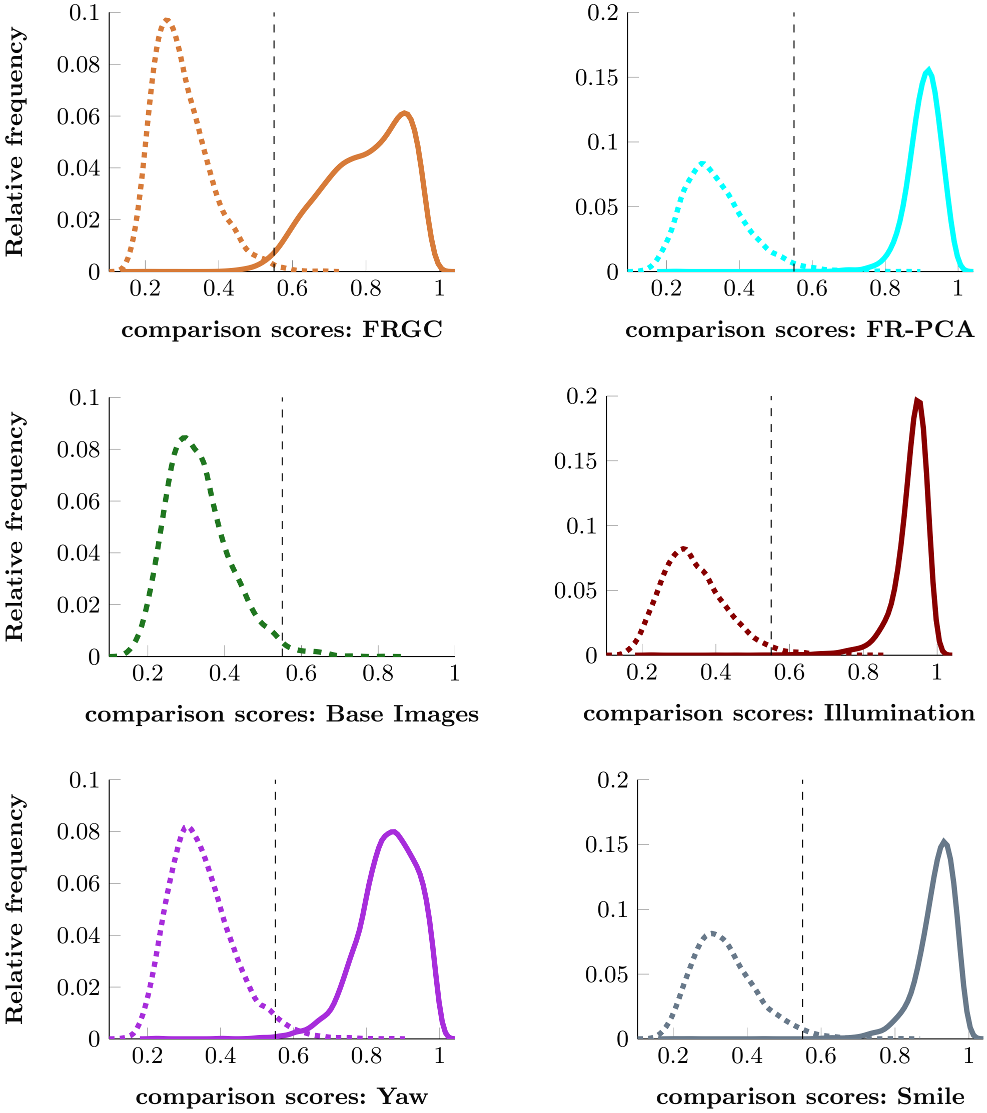


Fig. 7: Mated and non-mated comparison scores computed with VGGFace2 [24]. Thick solid line represents the kernel density curve of mated comparison scores, Thick dotted line represents the non-mated comparison scores and black dashed line represents the threshold @ FMR= 0.1% on LFW [12] dataset. Note that our base image dataset includes a single image per identity, therefore only depicting the non-mated comparison score distribution.

6 Conclusion and Future Work

To solve the privacy-related issue with real datasets and overcome the shortage of training data, we introduce PCA-FR-Guided sampling for generating mated samples in a non-deterministic manner. Unlike controlled face image editing techniques operating in the latent space, we apply PCA to find semantically meaningful directions. While moving latent vectors into these directions, the identity of the underlying face image is preserved by progressive supervision with a pre-trained face recognition system. With the newly created Synthetic Mated samples dataset (SymFace Dataset) with 77,034 images, we have evaluated state-of-the-art face quality assessment algorithms and biometric comparison score analysis to validate the applicability of the proposed approach. The well-separated distributions between mated and non-mated comparison scores indicate that synthetic mated samples generated with PCA-FR-Guided sampling are well suited for biometric performance tests. Furthermore, the analysis of face quality and the comparison scores is comparable to observations made in real datasets, indicating the usefulness of the proposed approach.

Although this work has illustrated to include synthetic samples in face recognition performance tests, we emphasise the open challenge to mimic the full extent of intra-identity variation measurable in bona fide datasets. Future works should also focus on an exploratory analysis of the different principal components, thereby exploring the latent space of StyleGAN and strengthening the understanding of the internal data representation. We foresee using the proposed approach to reduce the need for large training sets and minimise the demographic bias by diversifying latent space in synthetic generation schemes.

Appendix

Datasets		PCA-FR	Illumination Quality	Smile	Yaw
FRGC v2.0	SER-FIQ	1.17	1.19	1.13	3.25
	FaceQnet v1	0.11	0.12	0.11	0.13
Base Images	SER-FIQ	0.02	0.01	0.01	0.27
	FaceQnet v1	0.01	0.01	0.01	0.01

Table 2: KL-Divergences between quality score distributions in Figure 5.

Datasets		PCA-FR	Illumination Quality	Smile	Yaw
FRGC v2.0	Mated	0.42	0.79	0.17	0.72
	Non-mated	0.27	0.28	0.32	0.33
Base Images	Mated	/	/	/	/
	Non-mated	0.01	0.20	0.02	0.01

Table 3: KL-Divergences between comparison score distributions in Figure 7.

References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7617–7627 (2021)
2. Colbois, L., Pereira, T.d.F., Marcel, S.: On the use of automatically generated synthetic image datasets for benchmarking face recognition. arXiv preprint arXiv:2106.04215 (2021)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
4. Council of European Union: Council regulation (EU) no 2226/2017: Establishing an Entry/Exit System (EES) (2017), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R2226
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
6. Frontex: Best practice technical guidelines for Automated Border Control (ABC) systems (2015)
7. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and vision computing **28**(5), 807–813 (2010)
8. Grother, P., Ngan, M.L., Hanaoka, K.K.: Ongoing face recognition vendor test (frvt) part 2: Identification. NIST Interagency Report (2018)
9. H. Zhang, M. Grimmer, R.R., K. Raja, C.B.: On the applicability of synthetic data for face recognition. In: IEEE Intl. Workshop on Biometrics and Forensics. pp. 1–6 (2021). https://doi.org/10.1109/IWBF50991.2021.9465085
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hernandez-Ortega, J., Galbally, J., Fierrez, J., Beslay, L.: Biometric quality: Review and application to face recognition with faceqnet. arXiv preprint arXiv:2006.03298 (2020)

12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments (October 2007)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proc. of the IEEE Intl. Conf. on Computer Vision. pp. 1501–1510 (2017)
14. International Civil Aviation Organization: Machine readable passports – part 9 – deployment of biometric identification and electronic storage of data in emrtds (2015), https://www.icao.int/publications/pages/publication.aspx?docnum=9303
15. International Organization for Standardization: ISO/IEC TR 29794-5:2010 Information technology — Biometric sample quality — Part 5: Face image data (2010)
16. International Organization for Standardization: ISO/IEC 29794-1:2016 Information technology — Biometric sample quality — Part 1: Framework (2016)
17. International Organization for Standardization: ISO/IEC 2382-37:2017 Information technology — Vocabulary — Part 37: Biometrics (2017)
18. International Organization for Standardization: ISO/IEC WD 29794-1:2021 Information technology — Biometric sample quality — Part 1: Framework (2021)
19. International Organization for Standardization: ISO/IEC WD 29794-5:2021 Biometric Sample Quality: Face image data: Biometrics (2021)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
22. Pearson, K.: On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science **2**(11), 559–572 (1901)
23. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Jin Chang, Hoffman, K., Marques, J., Jaesik Min, Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). pp. 947–954 vol. 1 (June 2005)
24. Qiong, C., Li, S., Weidi, X., Omkar, P., Andrew, Z.: Vggface2: A dataset for recognising faces across pose and age (2018)
25. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE transactions on pattern analysis and machine intelligence (2020)
26. Terhorst, P., Kolf, J., Damer, N., Kirchbuchner, F., Kuijper, A.: Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 5651–5660 (2020)
27. Zhang, C., Liu, S., Xu, X., Zhu, C.: C3ae: Exploring the limits of compact model for age estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12587–12596 (2019)

Searching for APN functions by polynomial expansion

Maren Hestad Aleksandersen[1], Lilya Budaghyan[1], and Nikolay Stoyanov Kaleyski[1]

Department of Informatics, University of Bergen
lilya.budaghyan@uib.no, maren.aleksandersen@gmail.com,
nikolay.kaleyski@uib.no

Abstract. We investigate how far the approach of searching for APN functions by expanding their univariate representation can be pushed. We present some theoretical tricks that can be used to speed up the search up to EA-equivalence. We conduct systematic experiments over $\mathbb{F}_{2^8}$ and $\mathbb{F}_{2^9}$ and partition the resulting functions using the differential spectrum of their orthoderivatives. We find one new APN instance over $\mathbb{F}_{2^8}$. We also find 15 APN instances over $\mathbb{F}_{2^8}$ and 19 APN instances over $\mathbb{F}_{2^9}$ that are CCZ-inequivalent to the known infinite APN families. We see that they have differential spectra corresponding to known APN instances, but observe that the representatives that we obtain are significantly simpler than the known ones. We thus conclude that polynomial expansion deserves to be investigated in more detail.

Keywords: APN functions · differential uniformity · polynomial expansion.

1 Introduction

An (n, n)-function is a mapping with n input bits and n output bits. Nonlinear components of block ciphers are typically modeled as (n, n)-functions, and their properties are crucial for the security of the ciphers. One of the most powerful known attacks is differential cryptanalysis [2]. The best resistance to it is provided by APN (almost perfect nonlinear) functions which also have many other connections to mathematics and computer science (see e.g. [6] for a comprehensive survey).

Finding APN functions is difficult and many computational procedures have been developed, e.g. [1], [3], [11]. These typically exploit a representation or some property of the functions, and have produced thousands of CCZ-inequivalent APN instances. A disadvantage is that these procedures (and their implementation) can be complicated. A more serious drawback is that the obtained functions often have a very complicated form, which makes it difficult to e.g. generalize them into infinite constructions.

Some of the earliest known polynomial APN functions (e.g. [7] or [8]) were found by polynomial expansion. This amounts to adding terms to the polynomial representation of a function F, and checking whether the resulting functions are APN. This method is easy to implement, and produces functions with a simple representation upon success. Despite this, it has not been considered seriously in the literature. In particular, no well-documented results exist showing how far it can be taken, and what searches have been performed.

In this abstract, we report on our computational results of applying polynomial expansion to some known APN functions over $\mathbb{F}_{2^8}$ and $\mathbb{F}_{2^9}$. We obtain many APN functions in this way, and use the differential spectra of the orthoderivatives [5] to partition them into classes. We find 16 classes in $\mathbb{F}_{2^8}$ and 19 classes in $\mathbb{F}_{2^9}$ of APN functions that are CCZ-inequivalent to representatives from the known infinite families. In the case of $\mathbb{F}_{2^8}$, one of the classes is completely new. The remaining classes match those of known APN instances (found by e.g. the method from [1] or [11]), but our representatives have a significantly simpler representation (for instance, only 5 instead of 44 terms). When the initial function that we expand is a monomial, we introduce some theoretical tricks that can be used to restrict the choice of coefficients (up to EA-equivalence) and significantly speed up the search.

We thus conclude that the polynomial expansion approach can still produce useful results, and deserves to receive more attention that it currently does.

2 Background and notation

An (n, n)**-function**, or vectorial Boolean function, is a map from the finite field $\mathbb{F}_{2^n}$ to itself. Any (n, n)-function can be uniquely represented as a univariate polynomial of the form $F(x) = \sum_{i=0}^{2^n-1} a_i x^i$ for $a_i \in \mathbb{F}_{2^n}$. The largest binary weight of an exponent i with $a_i \neq 0$ is called the **algebraic degree** of F, denoted $\deg(F)$. If $\deg(F) \leq 1$, we say that F is **affine**, and if $\deg(F) = 2$, we say that F is **quadratic**. An affine F with $F(0) = 0$ is called **linear**.

For an (n, n)-function F, we denote by $\delta_F(a, b)$ the number of solutions $x \in \mathbb{F}_{2^n}$ to $F(a+x)+F(x) = b$ for $a \in \mathbb{F}_{2^n}$ and $b \in \mathbb{F}_{2^n}$. The **differential uniformity** of F is $\delta_F = \max_{a,b \in \mathbb{F}_{2^n}, a \neq 0} \delta_F(a, b)$. The lower the value of δ_F, the more resistant F is to differential attacks. Clearly, $\delta_F \geq 2$ for any (n, n)-function F. If $\delta_F = 2$, we say that F is **almost perfect nonlinear (APN)**.

Two (n, n)-functions F and G are **CCZ-equivalent** (Carlet-Charpin-Zinoviev-equivalent) if there exists an affine permutation A of $\mathbb{F}_{2^n} \times \mathbb{F}_{2^n}$ mapping the graph $\{(x, F(x)) : x \in \mathbb{F}_{2^n}\}$ of F to the graph of G. CCZ-equivalence is the most general known relation preserving APN-ness, and so APN functions are typically classified up to CCZ-equivalence. A special case of CCZ-equivalence is EA-equivalence. We say that F and G are **EA-equivalent** (extended affine equivalent) if $A_1 \circ F \circ A_2 + A = G$ for some affine A_1, A_2, A with A_1 and A_2 being permutations. Two quadratic APN functions are CCZ-equivalent if and only if they are EA-equivalent [9]. Furthermore, most of the known APN functions are quadratic, or CCZ-equivalent to quadratic (see e.g. [6] for a general survey, or [4] for a survey of the known infinite constructions). Thus, in practice, it is often enough to test EA-equivalence.

The **orthoderivative** π_F is a function uniquely associated with a quadratic APN function F. If F and G are EA-equivalent (and hence also CCZ-equivalent), then so are π_F and π_G [5]. The multiset of the values of $\delta_F(a,b)$ through all $a, b \in \mathbb{F}_{2^n}$ is called the **differential spectrum** of F, and is invariant under EA-equivalence. The differential spectra of the orthoderivatives are a very strong invariant for quadratic APN functions that has almost the same distinguishing power as an EA-equivalence test [5].

3 Polynomial expansion

Consider an initial (n,n)-function F. We conduct an exhaustive search over all functions of the form $F + c_1 x^{i_1} + \cdots + c_K x^{i_K}$ for a natural number K and all possible coefficients c_j and exponents i_j. We restrict the exponents to quadratic ones since we use the orthoderivatives to distinguish between inequivalent functions, and they are only defined for quadratic APN functions; furthermore, most of the known APN functions are quadratic, so the probability of finding non-quadratic ones in this way is very low. For small values of K, we consider all coefficients $c_j \in \mathbb{F}_{2^n}$. When the search becomes too slow, we restrict the coefficients to a subfield of $\mathbb{F}_{2^n}$. We do not perform searches with coefficients restricted to $\mathbb{F}_2$ since all quadratic APN functions with binary coefficients over $\mathbb{F}_{2^n}$ have been classified for $n \leq 9$ [10].

When F is a monomial, we can speed up the search as follows. Let $F(x) = x^d$ and $G(x) = x^d + cx^i$. Composing $L_1 \circ F \circ L_2$ with $L_1(x) = x/a^d$ and $L_2(x) = ax$ for $0 \neq a \in \mathbb{F}_{2^n}$, we obtain the EA-equivalent $G'(x) = x^d + ca^{i-d}x^i$. Thus, the coefficient of the first expansion term can be multiplied by a^{i-d} for $0 \neq a$. Thus, having tried c, we can ignore all coefficients of the form ca^{i-d} for $0 \neq a \in \mathbb{F}_{2^n}$. Similarly, we can take $L_1(x) = x^2$ and $L_2(x) = x^{2^{n-1}}$, and obtain the EA-equivalent $G''(x) = x^d + c^2 x^i$. Thus, c can be raised to any power of 2. Restricting the coefficient of the first term in this way significantly reduces the search space, and allows us to perform e.g. searches with $K = 5$ terms in $\mathbb{F}_{2^9}$ when the initial function is a monomial, while in the case of polynomials, we must restrict ourselves to $K = 4$ terms due to long running times.

4 Computational results

We consider as an initial function F a single representative from each CCZ-class represented by the quadratic infinite APN families. In the case of $n = 8$, we run searches with coefficients in $\mathbb{F}_{2^8}$ for up to 3 terms when F is a monomial, and up to 2 terms otherwise. Restricting the coefficients to $\mathbb{F}_{2^4}$, we attempt to add 4 terms, and with coefficients in $\mathbb{F}_{2^2}$ we are able to go up to 6 terms. All running times are within 100 hours; pushing the search further may be possible, but would require considerable computational effort.

We find 16 classes (according to the orthoderivative's differential spectrum) CCZ-inequivalent to the known infinite families. Among these, the function $x^3 + \beta x^{18} + \beta x^{66} + \beta^2 x^{132}$ (where β is primitive in $\mathbb{F}_{2^2}$) is completely new, having differential spectrum of the orthoderivative $0^{38196}, 2^{22008}, 4^{4608}, 6^{456}, 8^{12}$ (with the multiplicity of each element written in superscript). The remaining 15 differential spectra match those of known APN instances; however, our representations are significantly shorter and better structured than the known ones in many cases. For instance, one of our representatives, viz. $x^5 + x^9 + \beta x^{17} + \beta x^{65} + \beta^2 x^{170} x^{80} + \beta x^{96} + x^{144}$, has 7 terms, with coefficients in $\mathbb{F}_{2^2}$. It is equivalent to a known instances obtained in [1] having 36 terms with various coefficients.

For $\mathbb{F}_{2^9}$, the situation is similar. For coefficients in $\mathbb{F}_{2^9}$, we go up to 2 terms, and then restrict the coefficients to $\mathbb{F}_{2^3}$. We can go up to 5 terms for monomials (using the simplification described above) and up to 4 terms when the initial function is a polynomial. The running times are within 600 hours. We find 19 orthoderivative differential spectra that are not represented by the known infinite APN families. All of them correspond to known instances but some of our representatives are significantly simpler. For instance, one of the functions in [1] has 44 terms, while our representative can be written as $x^3 + \gamma x^{10} + x^{17} + \gamma x^{66} + x^{80}$, with γ primitive in $\mathbb{F}_{2^3}$.

Due to space limitations, we do not provide a full list of the representatives that we find here; one is available in the first author's master thesis, or online at `https://boolean.h.uib.no/mediawiki/index.php/APN_functions_obtained_via_polynomial_expansion_in_small_dimensions`.

References

1. Beierle C, Leander G. New instances of quadratic APN functions. arXiv preprint arXiv:2009.07204. 2020 Sep 15.
2. Biham E, Shamir A. Differential cryptanalysis of DES-like cryptosystems. Journal of CRYPTOLOGY. 1991 Jan 1;4(1):3-72.
3. Budaghyan L, Calderini M, Carlet C, Coulter R, Villa I. Generalized isotopic shift construction for APN functions. Designs, Codes and Cryptography. 2021 Jan;89(1):19-32.
4. Calderini M, Budaghyan L, Carlet C. On known constructions of APN and AB functions and their relation to each other. Rad Hrvatske akademije znanosti i umjetnosti: Matematike znanosti. 2021 Aug 25(546= 25):79-105.
5. Canteaut A, Couvreur A, Perrin L. Recovering or Testing Extended-Affine Equivalence. arXiv preprint arXiv:2103.00078. 2021 Feb 26.
6. Carlet C. Boolean functions for cryptography and coding theory. Cambridge University Press, 2021.
7. Dillon JF. APN polynomials and related codes. InBanff Conference, Nov. 2006 2006.
8. Edel Y, Kyureghyan G, Pott A. A new APN function which is not equivalent to a power mapping. IEEE Transactions on Information Theory. 2006 Jan 23;52(2):744-7.
9. Yoshiara S. Equivalences of quadratic APN functions. Journal of Algebraic Combinatorics. 2012 May;35(3):461-75.
10. Yu Y, Kaleyski N, Budaghyan L, Li Y. Classification of quadratic APN functions with coefficients in F2 for dimensions up to 9. Finite Fields and Their Applications. 2020 Dec 1;68:101733.
11. Yu Y, Wang M, Li Y. A matrix approach for constructing quadratic APN functions. Designs, codes and cryptography. 2014 Nov;73(2):587-600.

Stupid, Evil, or Both?
Understanding the Smittestopp conflict

Hans Heum[1]

Simula UiB
Merkantilen (3rd floor)
Thormøhlensgate 53D
N-5006 Bergen, Norway.
hansh@simula.no

Abstract. Like many governments, the Norwegian government provided a contact tracing application to help in combating the COVID-19 pandemic at its outset. However, the application was widely criticized for enabling an unacceptable intrusion into its subjects' lives, leading to its discontinuation only four months into the pandemic. In this essay, we will take a closer look at what went wrong, attempt to gain a deeper understanding of the passionate nature of the conflict, and how both sides came to view the other as being either stupid, or evil, or both.[1]

Keywords: Privacy · Contact Tracing · Code of Ethics · Smittestopp

1 Introduction

On March 31, 2020, the preprint *Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing* [4] appeared online, having been accepted for publication in the prestigious journal Science.

On April 16, Norwegian authorities released their solution: Smittestopp. Developed by the Norwegian Institute of Public Health (NIPH),[2] in collaboration with Simula Research Laboratory, the app had the dual purposes of providing automatic, digital contact tracing, and of gathering data to monitor the spread of the coronavirus [13].

On June 2, Amnesty International presented an open letter to the Norwegian Minister of Justice and Public Security, claiming that the app was violating fundamental human rights, and urging the Norwegian government to immediately roll back its employment [16].

On June 12, the Norwegian Data Protection Authority (NDPA)[3] announced a temporary ban on the processing of any and all data gathered by the app [24].

On June 15, the app was discontinued [11].

To this day, all parties appear convinced of their moral high ground, viewing the opposing side as "either stupid, or evil, or both".[4] On the one side stands NIPH, together with the developers of Simula Research Laboratories, apparently firm in their belief that the loss of privacy was a small price to pay for data on the spreading of the disease [18]. On the other side stands NDPA and Amnesty International, firm in their belief that such an intrusion would violate the Universal Declaration of Human Rights.[5]

How did such a fundamental disagreement arise? Is there hope for reconsiliation? And, finally, who are right? *Should* Smittestopp's intrusive data gathering capabilities have been accepted in the name of science and public health?

2 A Doctor's Oath

In attempting to understand the viewpoints of the opposing parties, it is crucial to try to understand the respective research communities surrounding them. A useful tool towards this goal is to look at how *codes of ethics* appear in each: this can help reveal common biases brought on through years of training and engaging with the community—even if those same codes of ethics are far from at the front of a researcher's mind in the middle of a heated argument.

[1] The following essay was originally written October 2020; it has been updated for publication in 2021.

[2] Norwegian: Folkehelseinstituttet (FHI).

[3] Norwegian: Datatilsynet.

[4] Quoting a source close to the project, who for the purpose of this essay shall remain anonymous, in conversation with the author.

[5] Specifically, Article 12; see [2].

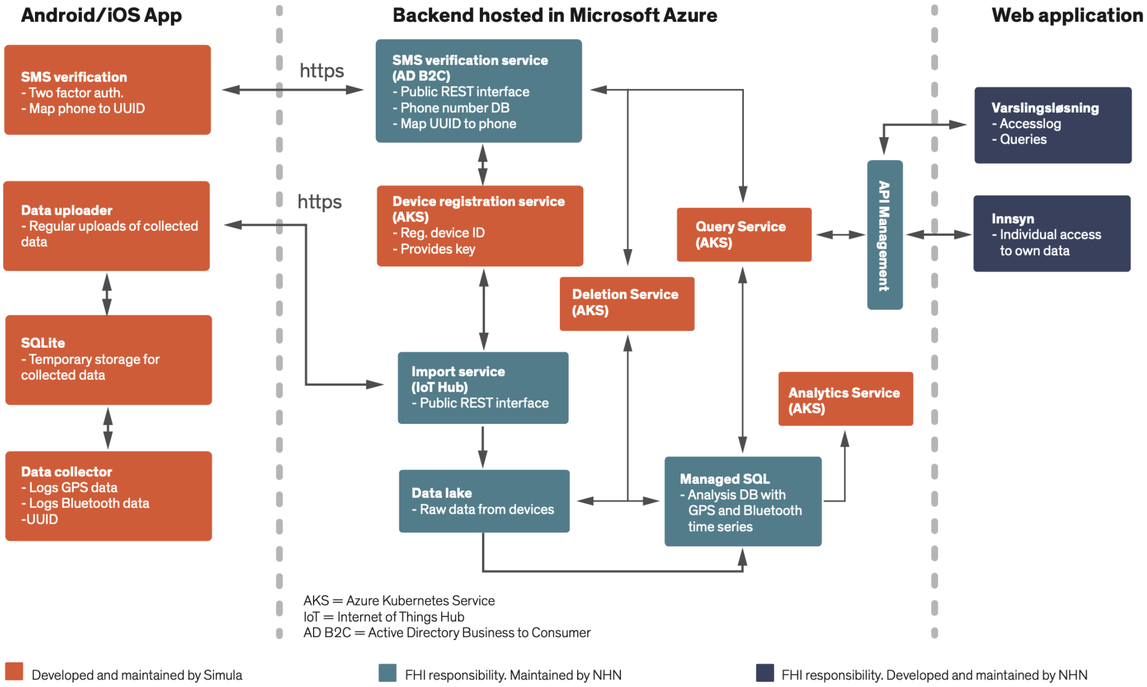


Fig. 1. An overview of the application architecture of Smittestopp. As can be seen (bottom-left), the app collects GPS data and Bluetooth data, and connects them with a unique identifier (UUID). Additionally, it collects the users phone number, as needed for the two-factor authentication, and associates it with their UUID (upper-left). Note that the UUID remains constant for each user. It seems likely that, for someone with full access to the servers, connecting parts of the data set to specific users would be a trivial task. Figure taken from [20, page 11]. The servers were located in Ireland, and were operated by Microsoft.

To start, it is commonly taught that Hippocrates wrote the first binding treatise on medical ethics. Hailing from between the 5th and the 3rd centuries BC, it included an oath to be sworn by all practitioners of medicine [3]. The oath spells out guidelines relating to the sharing of medical knowledge, and to the treatment and care of patients, with the central theme being that patients should be regarded as human beings rather than scientific subjects.

The oath survives today in modernized forms, and while several versions exist, the most widely adopted modern version of the oath is the one written in 1964 by Dr. Louis Lasagna, Academic Dean at Tufts University School of Medicine [17]. This is the version that will be quoted here.

One line of the oath reads as follows:

> *I will prevent disease whenever I can, for prevention is preferable to cure.*

Many members of the Norwegian Institute of Public Health will likely have sworn this oath upon graduation. If they see the potential for disease prevention, then they will quite literally have a moral obligation to pursue it, as long as it doesn't otherwise contradict the oath, or the law at large.

The oath is not silent on the subject of privacy either, stating:

> *I will respect the privacy of my patients, for their problems are not disclosed to me that the world may know.*

Meanwhile, in January 2020, the Nuffield Council On Bioethics released their report, "Research in global health emergencies: ethical issues" [15]. Endorsed by the World Health Organization [28], the report has become a guiding document to research on the pandemic. Chapter 9 of the report deals with the handling of data. To quote from the chapter summary (page 186):

> *Sharing data and samples between humanitarian actors, or for future research use, can play an important role in helping reduce suffering in many ways, both during emergencies and in the routine surveillance that forms part of emergency preparedness. However, sharing may also bring with it risks of harm and exploitation* [. . .] *Sharing is vital for effective research collaboration, but it must not be exploitative.*

Reading this, one is reminded of the first promise of the Hippocratic oath:

> *I will respect the hard-won scientific gains of those physicians in whose steps I walk, and gladly share such knowledge as is mine with those who are to follow.*

In other words, they are talking about the more-or-less open sharing of potentially private or harmful data used in research. It should be obvious to anyone that such sharing is to be done with great care, and the main tool to guard against this kind of misconduct, is oversight.

It was likely quite clear to the medical professionals involved with the design of Smittestopp that there would be no danger of such misconduct—after all, they were looking to publish *statistics*, not the names and phone numbers of their subjects. As is common practice, such data would be purged from the data set before analysis even began. Meanwhile, the data itself would remain securely hidden away on servers in Ireland, guarded against intrusion at all hours, and the scientists involved would have careful oversight on the manners in which the data was used in their research.

With a moral obligation to prevent the spreading of the disease, and with the collection of data seen as vital to this effort, together with the fact that no disclosure of personal information was to take place, one begins to understand how collecting movement data on infected individuals was seen as an absolutely vital part of the application's purpose and design, and it's likely that the medical researchers didn't see the presence of an ethical dilemma here at all. On the contrary, one imagines that they must have been genuinely confused at the backlash that was to follow.

3 A Human Right

Article 12 of the Universal Declaration of Human Rights [2] reads:

> *No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.*

The extents and limits of this right are (as with most legal texts) subject to fervent debate. Nevertheless, it may well be viewed as foundational for the *privacy community*.

To quote Eric Hughes' influential 1993 text, *A Cypherpunk's Manifesto* [6]:

> *We cannot expect governments, corporations, or other large, faceless organizations to grant us privacy out of their beneficence. It is to their advantage to speak of us, and we should expect that they will speak. To try to prevent their speech is to fight against the realities of information.* [. . .] *We must defend our own privacy if we expect to have any.*

Their greatest fear is typically an Orwellian nightmare, in which individuals are monitored and tracked, free movement and speech are distant memories, and even free thought is under constant attack. Furthermore, they view the road to such a society as a slippery slope towards which all authority is inevitably drawn.[6] Therefore, one must be vigilant and nip in the bud any and all attempts by people in power to relax privacy rights or increase surveillance efforts. Needless to say, they view all data collection and monitoring with extreme skepticism, regardless of its stated purpose.

On the other hand, in the modern world, it is not only state actors that infringe upon our privacy. Modern communication technology has made it possible for companies like Google and Facebook to build entire industries around harvesting personal data and selling it to advertising firms. While their motivations may seem less sinister than those envisioned of the state, the power they amass through such data harvesting[7] cannot be overstated. Such power erodes democracy itself, as can be seen by the fact that governments now regularly plead with the leaders of big international corporations, as if they were sovereign entities themselves.[8]

For the above reasons, many members of the privacy community will oppose *any* effort to implement digital contact tracing—even those treating the privacy of its subjects with the utmost care and respect. On the other hand, there are many who would agree that if an application could be designed such that it is verifiably impossible (or at least verifiably very difficult) for anyone, *including the maintainers of the system*, to exploit the system for anything beyond contact tracing, *then* the benefits of digital contact tracing will outweigh the tiny loss of privacy.

[6] And there is certainly historical precedent for such a view! As the saying goes, "Power corrupts; absolute power corrupts absolutely."

[7] Not to mention the power gained through the billions earned in targeted advertising.

[8] The subject of digital contact tracing actually provides a very clear example of this, as we shall see later.

How to achieve fully privacy-preserving digital contact tracing remains an open problem, and no widely-agreed-upon solution exists [25], but at a minimum, the application should have an open source code, and a description of all data processing should be available and clearly presented, in order that any interested party may verify the design guarantees.[9] This, and more, is spelled out in a Joint Statement on Contact Tracing, dated April 19, 2020, and signed by hundreds of researchers from around the world [7].

But with Smittestopp *explicitly* designed to collect and exploit data beyond the purpose of contact tracing, storing both personal information and movement data on central servers with no clear separation between the two, *and* launching with a closed source code, the app simply failed on all counts. It should not be difficult to see how the privacy community was both shocked and enraged by its design. To the privacy enthusiast, the infringements represented by Smittestopp were simply unacceptable.[10]

4 Late Lessons, Early Warnings

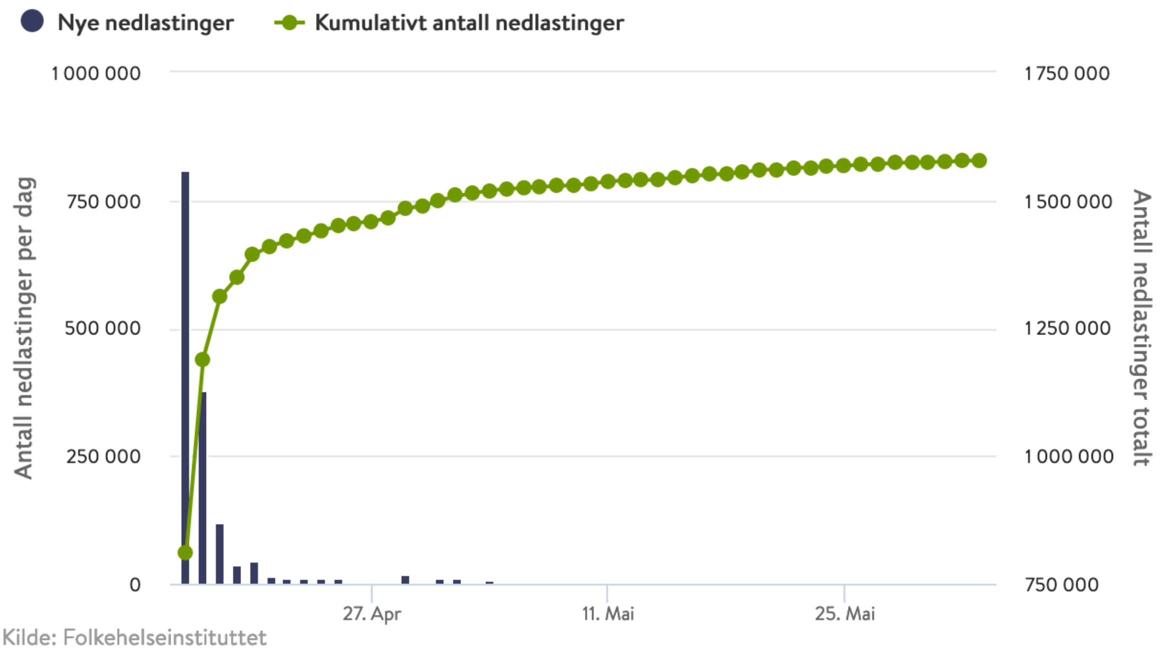


Fig. 2. A graph showing the downloads of Smittestopp over time. Each bar represents one day of downloads. The dotted line represents total downloads. Taken from [10].

In total, roughly 30% of the Norwegian population downloaded the app, with the majority of downloads happening the first few days after launch. In the following weeks, the number of users quickly flattened, and then started falling. By June, the number of active users had dropped to less than 14% among Norwegians 16 years or older.[11]

[9] This is usually done by trying to come up with ways to *break* them, and communicating to the maintainers any security holes they find. Trust in the system as a whole then comes from a shared trust in its maintainers to continually roll out security patches, and in the community to vigilantly find clever new ways to attack it.

[10] There are further issues with the Smittestopp design beyond those discussed here; one is the use of static identifiers. In a blog post, Ian Levy, the technical director at the National Cyber Security Centre UK (NCSC), writes that “In any contact tracing app that broadcasts something to be picked up by others, there are risks. There are a range of schemes from having a fixed-for-all-time ID that’s constantly broadcast (which would be silly as anyone can see if you’re around), through to schemes that make it exceptionally difficult to work out what’s going on. There are a set of well known attacks that all apps have to mitigate.” [8] The above paranthesis turns particularly bemusing once one realizes that this is exactly what Smittestopp did, as can be seen in Fig. 1.

[11] These numbers were calculated with user data from NIPH [10] and population data from Statistics Norway [21]. Note that the first number is an estimate, due to the fact that re-downloads of the app by the same person cannot be precisely accounted for.

It is fair to assume that the dwindling user base can be blamed on the large number of highly critical articles published in Norwegian news media in the weeks following Smittestopp's launch, [12] together with the authorities' inability or unwillingness to answer the worries of the populace. The small user base, the complete lack of public trust in the application, and the low infection numbers in Norway at the time, were all factors leading to NDPA's temporary ban, as announced on June 12 [24]. In total, less then ten warnings of close contact was issued by the app to an end user [19]. Furthermore, according to a source close to the project, none of the data collected ended up being processed for research before the ban was issued and the data deleted. While this may be comforting to privacy enthusiasts, it also means that, by failing to gain the public's trust, the NIPH failed to achieve *any* of their goals with the application.

What could they have done differently? First, and foremost, they could have chosen to pay attention to—and engage in—the international discourse and developments on privacy preserving digital contact tracing that was happening at the time.

On March 19, a full month before the launch of Smittestopp, the European Data Protection Board (EDPB) released a statement on the processing of personal data in context of the pandemic [23]. The statement included the following:

> *Personal data that is necessary to attain the objectives pursued should be processed for specified and explicit purposes.* [. . .] *The least intrusive solutions should always be preferred, taking into account the specific purpose to be achieved.*

The previously mentioned Joint Statement on Contact Tracing was released on April 19, shortly after Smittestopp's launch, and on April 21, EDPB published guidelines for digital contact tracing, to be adopted throughout the EEA. They include [22, page 7]:

> *In the context of a contact tracing application, careful consideration should be given to the principle of data minimisation and data protection by design and by default:*
> – *contact tracing apps do not require tracking the location of individual users. Instead, proximity data should be used;*
> – *as contact tracing applications can function without direct identification of individuals, appropriate measures should be put in place to prevent re-identification;*
> – *the collected information should reside on the terminal equipment of the user and only the relevant information should be collected when absolutely necessary.*

Secondly, they could have realized from the start that, codes of ethics aside, nothing will be achieved without the trust of the populace: without widespread adoption, digital contact tracing is simply not effective. This was already realized by the authors of the Science-published article that originally suggested digital contact tracing as a tool to combat COVID-19 [4, page 5]:

> *Successful and appropriate use of the app relies on it commanding well-founded public trust and confidence. This applies to the use of the app itself and of the data gathered.*

At this point, it should have been clear to the NIPH that development of a new version of Smittestopp should commence immediately—a version that neither relied on nor collected GPS data, that was in line with international regulation, and that aimed to earn the public's trust. Instead, they seem to have ignored the European legislation, and faced the criticism with an appeal to blind trust and "dugnadsånd",[13] with the implication being that anyone who refused to download the app was being selfish at the expense of those around them.

Then, on June 2, Amnesty International sent an open letter to Monica Mæland, then minister of justice and public security, in the process sparking the biggest media storm so far. The letter concludes [16]:

> *Amnesty International's Security lab identified the following features of the app that are highly concerning from a privacy perspective and do not comply with several human rights standards outlined above:*
> – *The app requires registration with a valid phone number. Thus, the operators of the app can tie any data upload to an identifiable individual.*
> – *The app collects GPS data. It stores a local copy, but also uploads this data to a central server. This allows operators of the app to track movement and location data of thousands of people who have the app installed. The Smittestopp app thus has the potential to be a mass surveillance tool.*

[12] While such articles are too numerous to list, see [5] for an early example.

[13] A Norwegian term characterising a neighbourly and helpful spirit, used repeatedly by then prime minister Erna Solberg at the outset of the pandemic.

> - *The app also uploads all user data to third-party Microsoft servers, which appear to be operating in Ireland.*
> - *The app also performs Bluetooth-based contact tracing. Apps running on devices in the proximity will exchange the respective unique identifiers and store them locally, along with a timestamp and signal strength. These records are also uploaded to the central server. This data is thus neither anonymised, nor decentralised, allowing app operators to track users' movements making it a privacy violation. Additionally, the use of unique identifiers could enable malicious actors to track users' movements using a distributed network of Bluetooth sensors. This is a privacy risk.*
>
> *Given the grave privacy risks to thousands of people, we wanted to alert you to this and urge you to immediately roll back the app in its current form and ensure that any contact tracing efforts are human rights respecting.*

As we've seen, the app was discontinued less than two weeks later.

5 Conclusion

In January 2018, The World Economic Forum published a Code of Ethics for Researchers [27]. The document attempts to define a set of principles that may be adopted by all who do research, irrespective of field of study. They are:

- *Engage with the public.*
- *Pursue the truth.*
- *Minimize harm.*
- *Engage with decision-makers.*
- *Support diversity.*
- *Be a mentor.*
- *Be accountable.*

If we are to judge the parties based on these "field-agnostic" principles, it seems clear that the NIPH failed on three points. First, they failed to engage with the public. Instead, they brushed aside complaints, and refused to accept criticism.

Second, in their "pursuit of truth", and in not realizing (or acknowledging) the importance and non-trivial nature of privacy preservation, they did not seek to minimize harm, even as the harmful nature of their practice was pointed out to them.

Third, as we've seen, they failed to engage with decision-makers, completely ignoring legislation on contact-tracing applications issude by the EU—legislation under which Norway, as a member of the EEU, was subject.

To their credit, though, once the NDPA issued their temporary ban on the processing of personal data in the app, the NIPH chose to immediately discontinue the app and delete all data gathered [11]; as such, they stood accountable.

On September 28, the Norwegian health minister announced that a new version of Smittestopp was in development [14], this time relying on the Google and Apple Exposure Notification system, or GAEN [1].[14] This architecture, based on one of the main proposals to come out of the Pan-European Privacy-Preserving Proximity Tracing organization, is decentralised, meaning that someone with access to the central servers will not learn anything about its user base, and it is designed from the bottom up with the privacy of its users in mind.[15]

Due to severe restrictions on the Bluetooth capabilities of third-party applications on iPhone, it is now generally agreed that for a decentralized contact-tracing app to be effective, it *must* be based on GAEN.[16] Governments have pleaded with Apple to lift the restrictions on apps issued by official authorities for the purposes of contact tracing, but Apple won't budge, saying effectively that GAEN is their solution, and they can take it or leave it. To quote Michael Veale, one of the designers of the architecture underpinning GAEN [26],

[14] In the time since this essay was written, the new application has been released, with very little fanfare and with no media storm in sight [12]. With the number of downloads now surpassing one million, and with close to 5000 having used the app to notify that they have been infected, it's safe to say that the GAEN-based Smittestopp has been a relative success. [9] (Further statistics, like the total number of users notified of close contact, are unavailable due to the privacy-preserving nature of the application.)

[15] Whether it achieves a satisfying level of privacy remains however disputed. This goes both for the underlying protocol itself [25], and for GAEN in particular, as Google will upload data—including both personal and location data—on all Google Play Store users at 20 minute intervals [20].

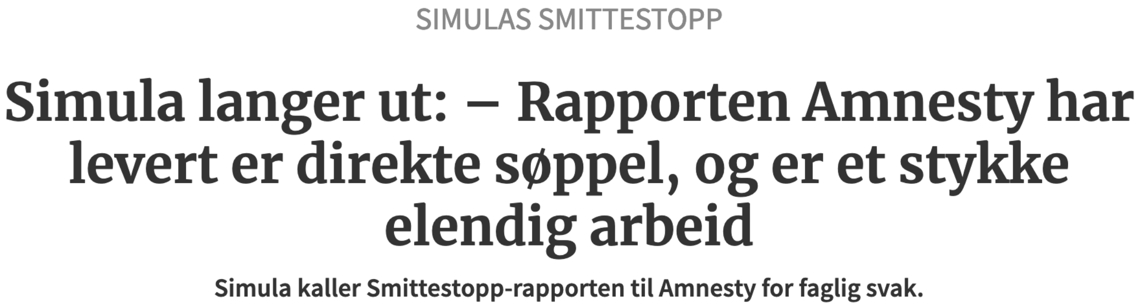
[16] This, dubbed "the Bluetooth problem", was overcome in the original Smittestopp design partly through the use of location data, and partly through the centralized design.

> [GAEN is] *great for individual privacy, but the kind of infrastructural power it enables should give us sleepless nights. Countries that expect to deal a mortal wound to tech giants by stopping them building data mountains are bulls charging at a red rag. In all the global crises, pandemics and social upheavals that may yet come, those in control of the computers, not those with the largest datasets, have the best visibility and the best—and perhaps the scariest—ability to change the world.* [...] *Law should be puncturing and distributing this power, and giving it to individuals, communities and, with appropriate and improved human-rights protections, to governments.*

In other words, pleading won't cut it.

6 Epilogue

Then, just as one might think that the dust had settled for, that the axes were finally buried and we were all ready to look ahead to a future of privacy-preserving digital contact tracing, Simula Research Laboratories, the developer of Smittestopp,[17] decided to speak up. [18]



SIMULAS SMITTESTOPP

Simula langer ut: – Rapporten Amnesty har levert er direkte søppel, og er et stykke elendig arbeid

Simula kaller Smittestopp-rapporten til Amnesty for faglig svak.

Fig. 3. The headline, taken from [18], roughly translates to *Simula lashes out: 'Amnesty's report is nothing but trash, and a lousy piece of work.'*

The statements of Fig. 3 are due to the corporate vice president of Simula, and were published on September 29, one day after the development of a new version was announced.[18] He goes on to say that Amnesty has exploited the trust they hold with the public, either for the purpose of bringing more attention to themselves, or to "further an activist agenda". Smittestopp, he claims, satisfied almost all of Amnesty's principles, with the "sole exception" being that it was gathering movement data on its users.

It is difficult to imagine what Simula was hoping to achieve with these statements. If nothing else, they witness a severe disconnect from the international community that has grown around constructing and analysing digital contact tracing. One would at the very least have thought they were *aware* of the many statements, guidelines, and legislative documents that were published since March 2020, particularly given the ban issued by the NDPA, but the corporate vice president's statements have cast even this into doubt. As a matter of fact, he seems to blame Amnesty for NDPA's ruling, implying that the NDPA issued a ban on Smittestopp simply because it was the popular thing to do at the time.

If it is true that researchers have a duty to interact with decision-makers and the public, then we can't help but wonder if there aren't better ways to do it.

[17] And, through company ownership, the author's employer.

[18] Simula chose not to partake in the development of the new version of Smittestopp.

References

1. Exposure notifications: Using technology to help public health authorities fight covid-19. https://www.google.com/covid19/exposurenotifications/. Accessed: 2020-10-07.
2. UN General Assembly. Universal declaration of human rights, 1948. https://www.un.org/en/universal-declaration-human-rights/index.html. Accessed: 2020-10-07.
3. Encyclopædia Britannica. Hippocratic oath, 2019. https://www.britannica.com/topic/Hippocratic-oath. Accessed: 2020-10-07.
4. Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, 368(6491), 2020.
5. Jonathan Simchai Hansen and Henrik Myhrvold Simensen. Lektor sendte e-post til høyskoleelever: Omtalte smittestopp-appen som 'farligere' enn coronaviruset. *VG*, April 2020. https://www.vg.no/nyheter/innenriks/i/GGLEw9/lektor-sendte-e-post-til-hoeyskoleelever-omtalte-smittestopp-appen-som-farligere-enn-coronaviruset. Accessed: 2020-10-07.
6. Eric Hughes. A cypherpunk's manifesto, March 1993. https://www.activism.net/cypherpunk/manifesto.html. Accessed: 2020-10-07.
7. Dali Kaafar et al. Joint statement on contact tracing, April 2020. https://www.esat.kuleuven.be/cosic/sites/contact-tracing-joint-statement/. Accessed: 2020-10-07.
8. Ian Levy. The security behind the nhs contact tracing app. https://www.ncsc.gov.uk/blog-post/security-behind-nhs-contact-tracing-app, May 2020. Accessed: 2021-01-07.
9. NIPH. Antall nedlastinger av, og antall meldt smittet gjennom, smittestopp. https://www.fhi.no/om/smittestopp/nokkeltall-fra-smittestopp/. Accessed: 2021-09-08.
10. NIPH. Antall nedlastinger og antall brukere av smittestopp. https://www.fhi.no/sv/smittsomme-sykdommer/corona/nokkeltall-fra-smittestopp/. Accessed: 2020-10-07.
11. NIPH. FHI stopper all innsamling av data i smittestopp. https://www.fhi.no/nyheter/2020/fhi-stopper-all-innsamling-av-data-i-smittestopp/. Accessed: 2020-10-07.
12. NIPH. Ny smittestopp-app klar for nedlasting. https://www.fhi.no/nyheter/2020/ny-smittestopp-app-klar-for-nedlasting/. Accessed: 2021-01-07.
13. NIPH. Smittestopp – ny app fra Folkehelseinstituttet. https://www.fhi.no/nyheter/2020/ny-app-fra-folkehelseinstituttet/. Accessed: 2020-10-07.
14. NIPH. Starter arbeid med ny løsning for digital smittesporing. https://www.fhi.no/nyheter/2020/fhi-stopper-all-innsamling-av-data-i-smittestopp/, September 2020. Accessed: 2020-10-07.
15. Nuffield Council on Bioethics. Research in global health emergencies, January 2020. https://www.nuffieldbioethics.org/publications/research-in-global-health-emergencies. Accessed: 2020-10-07.
16. Tanya O'Carroll and John Peder Egenæs. Concerns regarding the government of Norway's Smittestopp app, June 2020. https://www.digi.no/filer/Amnestys_brev_til_regjeringen_om_Smittestopp.pdf. Accessed: 2020-10-07.
17. Practo. The hippocratic oath: The original and revised version, March 2015. https://doctors.practo.com/the-hippocratic-oath-the-original-and-revised-version/. Accessed: 2020-10-07.
18. Martin Braathen Røise. Simula langer ut: – rapporten Amnesty har levert er direkte søppel, og er et stykke elendig arbeid. *digi*, September 2020. https://www.digi.no/artikler/simula-langer-ut-rapporten-de-har-levert-er-direkte-soppel-og-er-et-stykke-elendig-arbeid/500102?key=yJfvW2o8. Accessed: 2020-10-07.
19. Eystein Røssum. Færre enn ti har fått varsel frå smittestopp-appen. https://www.bt.no/nyheter/i/qL0MBO/faerre-enn-ti-har-faatt-varsel-fraa-smittestopp-appen, May 2020. Accessed: 2021-01-07.
20. Simula Research Laboratory and Simula Metropolitan. Sammenligning av alternative løsninger for digital smittesporing, September 2020. https://www.simula.no/sites/default/files/sammenligning_alternative_digital_smittesporing.pdf. Accessed: 2020-10-07.
21. SSB. Alders- og kjønnsfordeling i kommuner, fylker og hele landets befolkning 1986 - 2020. https://www.ssb.no/statbank/table/07459/. Accessed: 2020-10-07.
22. The European Data Protection Board. Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the covid-19 outbreak. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf, April 2020. Accessed: 2020-10-07.
23. The European Data Protection Board. Statement on the processing of personal data in the context of the covid-19 outbreak. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_statement_2020_processingpersonaldataandcovid-19_en.pdf, March 2020. Accessed: 2020-10-07.
24. Bjørn Erik Thon and Susanne Lie. Varsel om vedtak om midlertidig forbud mot å behandle personopplysninger – appen Smittestopp, June 2020. https://www.fhi.no/contentassets/7ac87ad803c3425688d6cc72e14924cf/20-02058-9-varsel-om-vedtak-om-midlertidig-forbud-mot-a-behandle-personopplysninger---appen-smittestopp.pdf. Accessed: 2020-10-07.
25. Serge Vaudenay. Centralized or decentralized? the contact tracing dilemma, May 2020. https://eprint.iacr.org/2020/531. Accessed: 2020-10-07.
26. Michael Veale. Privacy is not the problem with the apple-google contact-tracing toolkit. *The Guardian*, July 2020. https://www.theguardian.com/commentisfree/2020/jul/01/apple-google-contact-tracing-app-tech-giant-digital-rights. Accessed: 2020-10-07.

27. World Economic Forum Young Scientists. Code of ethics for researchers. wef.ch/coe, 2018. Accessed: 2020-10-07.

28. World Health Organization. Ethical standards for research during public health emergencies: Distilling existing guidance to support COVID-19 R&D, 2020. https://www.who.int/ethics/publications/ethical-standards-for-research-during-public-health-emergencies/en/. Accessed: 2020-10-07.

Machine Learning for Offensive Cyber Operations

Åvald Åslaugson Sommervoll[1][0000-0001-5232-5630],
Audun Jøsang[1][0000-0001-6337-2264]

University of Oslo, Problemveien 7, 0315 Oslo
aavalds@ifi.uio.no

Abstract. This paper gives a brief survey of existing and proposed applications of machine learning for offensive cyber operations, with particular emphasis on algorithmic cryptanalysis and penetration testing. For cryptanalysis at the algorithmic level, we cover attacks on historic ciphers as well as attacks on modern ciphers. For penetration testing, we cover works that have focused on defining structured attack approaches as well as some novel attacks where the potential merits need additional investigation.

Keywords: machine learning, offensive cyber operations, cryptanalysis, penetration testing, survey

1 Introduction

The arms race between cryptographers and cryptanalysts is an ancient one, with the earliest record of cryptanalysis dating back to the 9th century [14]. The attack described was frequency analysis effectively breaking the monoalphabetic substitution cipher; this implicated that for secure communication, the cryptographers would have to do something more advanced. A thousand years later the Germans used Enigma encryption, an encryption they thought to be unbreakable for communication during WWII. However, the huge joint effort of pre-WWII analysis of Polish mathematicians, paired with efforts from English and American scientists to develop cryptanalytical tools and methods, would show that it was indeed breakable [14]. Since WWII, Enigma encryption has been broken many times over because of its historical significance and as an effort to further offensive cyber operations[1] [10,17,13,12]. Some of these utilize machine learning techniques to speed up the attack [3,16]. Currently, in the arms race between cryptanalysts and cryptographers, it appears that cryptography has won, with standardized algorithms that are internationally recognized as secure. The arms race is far from over as new creative decryption attacks see light of day. However, since the algorithms themselves are deemed secure, modern attacks typically target the implementation, moving the hotspot of the current war from cryptology

[1] Note that we study offensive cyber operations: Testing and checking the integrity of existing cybersecurity defenses, not offensive cybersecurity: proactively predicting and removing threats in the system [1].

to cybersecurity[2]. There is a need for offensive cyber operations research to investigate the potential weaknesses and strengths of existing systems.

The rest of the paper is organized as follows: Section 2 involves a brief overview of machine learning and its impacts on cryptography. Section 3 covers some of the recent work on penetration testing using machine learning, in particular in terms of SQL injections. Finally, section 5 gives a brief concluding summary of this survey.

2 Cryptanalysis

Machine learning techniques are not easy to apply to the field of cryptoanalysis. This is because machine learning in general works by gradually inching closer to a good solution through *learning*, while modern crypto has many techniques that hide how close a cryptanalyst is to the solution; in other words obscuring learning. This obvious hurdle of machine learning in cryptoanalysis, may explain the rather short list of promising attempts using ML techniques. However, there has been documented some successes on classical systems such as Enigma [3,16]. Bagnall et al. cracked a two-rotor system of Enigma[3] which was based on using a genetic algorithm [3], but failing on 3 and 4 rotors. Sommervoll and Nilsen used the genetic algorithm to break the final step of Enigma decryption, finding all ten plugs of Enigma's plugboard faster than previous techniques [16]. More modern attacks are based on neuro-cryptanalysis first described by Dourlens in 1996 [6]. Since then, it has seen some limited success. Alani, in his neuro-cryptanalysis, attacks another classic but more modern cryptosystem DES and Triple-DES, with some success [2]. He does this by simulating the decryption under an unknown key using a neural network. In that, the input to his neural network are ciphertexts, and the output targets are the plaintexts. After training, he does not obtain the secret key, but ideally, a decryption machine that acts as the decryption algorithm with the key. He achieves an average bit accuracy of 91.7% for DES and 88.6% for Triple-DES. Also, in the field of neuro-cryptanalysis, a recent publication by Sommervoll in 2021 investigates the prospects of simulating an encryption algorithm as a neural network in what he refers to as the phantom gradient attack [15]. This attack does not draw from machine learning directly but attempts to use the same functions that train neural networks to train their way to the key. The trained network itself will, in this case, be uninteresting for prediction, but the trained weights will give the keys. Another example of neural-cryptanalysis is Aron Gohr's attack on Speck32/64 with deep learning [11]. Gohr did not use machine learning to recover the key directly, but used neural networks to distinguish between round reduced instances of Speck32/64 and random noise. He did this with great success, which is surprising from a cryptographic viewpoint. A recent follow-up paper by Benamira et al. investigates Gohr's findings [4]. They confirm his results, claim that his attack, while

[2] Side-channel attacks and espionage also have a rich history in humanity, though this history is so diverse that we do not cover it in this short review paper.

[3] Enigma encryption used had 3 to 4 rotors and a plugboard of 10 plugs during WWII.

impressive, is not really a novel cryptanalytical attack but is an optimization of the extraction of the low-data constrains.

3 Penetration testing

The field of penetration testing is considerably easier to unite with machine learning than algorithmic cryptanalysis. This is in large because machine learning agents can have the benefit of learning from humans, and the problems are not specifically designed to be difficult. Nonetheless, there is limited work done on automating the process of penetration testing with machine learning. Erdődi and Zennaro formalize part of this problem in the context of web hacking and reinforcement learning in [8]. The approach is called *Agent Web Model* that considers web hacking as a capture-the-flag (CTF) challenge. This model has seven layers of complexity, where layer 1 is the least complex, the agent is able to find links in objects, and layer 7 is the most complex; the agent is able to add files through a vulnerable object or create new database objects. In 2020 the authors demonstrated the potential of this approach by showing that reinforcement learning(RL) agents could solve CTF problems [18]. The authors showed that RL paired with techniques such as lazy loading, state aggregation, or imitation learning allowed the RL agent to perform more complicated tasks. Further, they argue that fully model-based agents may not be ideal as they are not as versatile; instead, they suggest model-free RL agents with rich a priori knowledge. Also, from 2020 is the work of Chaudhary *et al.* on automated post-breach penetration testing with RL [5]. The authors propose the idea of using RL agents to find sensitive files in a compromised network; however, from their paper, it seems that they are still working on obtaining specific results. Earlier work by Ghanem *et al.* compared a reinforcement learning agent called IAPTS (Automated Penetration Testing System) against blind automation and found that this RL agent performed better [9]. Their IAPTS agent has the possibility of human input on the decision policy; this will allow the agent to learn and better approximate the expert's decisions. Unfortunately, it does not yet perform all the tasks that a human expert is doing manually, but the authors indicate research directions to improve their approach. Some specific penetration testing tasks have seen very little research that utilizes offensive machine learning. To our knowledge, there is only one study for conducting SQL injections[4] [7]. Erdődi et al. simulate penetration testing in a capture-the-flag setting, where the agent can choose between a number of candidate SQL injection queries. From the queries, the agent learns to first find the correct escape before searching for the flag.

4 Conclusion

The literature on ML for offensive cyber operations is considerably smaller than the literature on ML for defensive cyber operations. In this review paper, we

[4] There are many machine learning papers for discovering SQL injection attacks.

reviewed studies that apply ML in offensive cyber operations. Algorithmic-level cryptanalysis seems to be challenging for ML because modern cryptographic algorithms are designed to make learning hard as there is no indication of close to correct decryptions. However, there are papers that document modest success on weak cryptosystems. Significant advances in this approach would be needed to facilitate more success against modern algorithms. Perhaps even less researched is to perform ML-based penetration testing. One reason for this could be because there are already many automated tools that cyber-ops professionals use and because it is very important that penetration tests are conducted properly. Because penetration testing is a vast field, and we are at a very early stage in research on applying ML for penetration testing, there seems to be a great potential for advances in this area. For example, in the area of SQL injection, which represents a significant part of penetration testing, we only identified one study on ML-based SQL penetration testing.

References

1. Aiyanyo, I.D., Samuel, H., Lim, H.: A systematic review of defensive and offensive cybersecurity with machine learning. Applied Sciences **10**(17) (2020). https://doi.org/10.3390/app10175811, https://www.mdpi.com/2076-3417/10/17/5811
2. Alani, M.M.: Neuro-cryptanalysis of des and triple-des. In: International Conference on Neural Information Processing. pp. 637–646. Springer (2012)
3. Bagnall, A.J., McKeown, G.P., Rayward-Smith, V.J.: The cryptanalysis of a three rotor machine using a genetic algorithm. In: ICGA. pp. 712–718 (1997)
4. Benamira, A., Gerault, D., Peyrin, T., Tan, Q.Q.: A deeper look at machine learning-based cryptanalysis. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. pp. 805–835. Springer (2021)
5. Chaudhary, S., O'Brien, A., Xu, S.: Automated post-breach penetration testing through reinforcement learning. In: 2020 IEEE Conference on Communications and Network Security (CNS). pp. 1–2. IEEE (2020)
6. Dourlens, S.: Applied neuro-cryptography and neuro-cryptanalysis of des. Master Thesis (1996). https://doi.org/10.13140/RG.2.2.35476.24960, advisor: Riesner, Christian
7. Erdodi, L., Sommervoll, Å.Å., Zennaro, F.M.: Simulating sql injection vulnerability exploitation using q-learning reinforcement learning agents. arXiv preprint arXiv:2101.03118 (2021)
8. Erdődi, L., Zennaro, F.M.: The agent web model: modeling web hacking for reinforcement learning. International Journal of Information Security pp. 1–17 (2021)
9. Ghanem, M.C., Chen, T.M.: Reinforcement learning for intelligent penetration testing. In: 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). pp. 185–192. IEEE (2018)
10. Gillogly, J.J.: Ciphertext-only cryptanalysis of enigma. Cryptologia **19**(4), 405–413 (1995)
11. Gohr, A.: Improving attacks on round-reduced speck32/64 using deep learning. In: Annual International Cryptology Conference. pp. 150–179. Springer (2019)
12. Lasry, G., Kopal, N., Wacker, A.: Cryptanalysis of enigma double indicators with hill climbing. Cryptologia pp. 1–26 (2019)

13. Ostwald, O., Weierud, F.: Modern breaking of enigma ciphertexts. Cryptologia **41**(5), 395–421 (2017)
14. Singh, S.: The code book: the science of secrecy from ancient Egypt to quantum cryptography. London: Fourth estate (2000)
15. Sommervoll, Å.: Dreaming of keys: Introducing the phantom gradient attack. In: 7th International Conference on Information Systems Security and Privacy, ICISSP 2021, 11 February 2021 through 13 February 2021. SciTePress (2021)
16. Sommervoll, Å.Å., Nilsen, L.: Genetic algorithm attack on enigma's plugboard. Cryptologia pp. 1–33 (2020)
17. Williams, H.: Applying statistical language recognition techniques in the ciphertext-only cryptanalysis of enigma. Cryptologia **24**(1), 4–17 (2000)
18. Zennaro, F.M., Erdodi, L.: Modeling penetration testing with reinforcement learning using capture-the-flag challenges and tabular q-learning. arXiv preprint arXiv:2005.12632 (2020)

are several such tools: Simuland[5] focuses on Microsoft Defence products, with a mapping to the Mitre ATT&CK Framework[6]; Cobalt Strike[7] provides a number of red team activities, from generating phishing emails to browser pivoting; Metasploit[8] supports the full attack scenario, from scanning for vulnerabilities to collecting credentials and generating a final report of the attack; PoshC2[9] focuses on post-exploitation and lateral movements with encrypted C&C traffic and features extensive logging of every action and response; Covenant[10] is a .NET C&C framework; Sliver[11] is a C&C red teaming tool; Atomic Red Team[12] is a library of simple detection tests mapped to the MITRE ATT&CK framework; and Merlin[13] is a popular post-exploitation C&C Tool.

None of these tools provide ground truth and a second challenge is correct labelling captured data with suitable granularity. One common approach, used e.g. by Garcia et al [3], is to label all data from malware infected machines as 'malicious' and everything else as either 'benign' or 'background'. This will entail that some benign data is erroneous labelled as malicious. A different approach is taken by Landauer et al [5], which combines knowledge of attack time with domain knowledge of the attack steps to label post-simulation – a similar approach is also taken by Buchanan et al [1]. Their approach does not target NIDS, and in addition, the labelling quality will depend on the domain knowledge and how it is implemented in the labelling process.

In our work, we instead build on the *DetGen*-tool by Clausen et al [2], where we can achieve finer grain control of the labelling compared with Garcia et al without the need for encoding domain expertise of each attack steps. Here, we encapsulate the malware in a container and label at the container-level, thus separating traffic arising from the malware from traffic arising from other processes in a machine. We extend [2] by encapsulating the Merlin C&C simulation tool in the container. Note that whilst Cobalt Strike was the most common tool used in malware in Recorded Future's 2021 report [7], it requires a licence. We therefore use Merlin, which is also a popular tool for simulating C&C.

2 An experiment using DetGen[Merlin] with Ghost

The Merlin C&C-framework has two main components: a server and a client. The server is configured to listen for HTTP-connections from the client, and sends C&C-commands to the client over this connection. The client software runs post-exploitation on a system you wish to control and will repeatedly connect to the server with a certain interval, also called a heartbeat. To avoid detection, the interval can be skewed to vary the interval.

The DetGen framework is built around Docker-Compose[14]. Each component in DetGen runs in a container, with a separate associated container to capture

[5] https://github.com/Azure/SimuLand

[6] https://attack.mitre.org/

[7] https://www.cobaltstrike.com

[8] https://www.metasploit.com/

[9] https://poshc2.readthedocs.io/en/latest/

[10] https://github.com/cobbr/Covenant

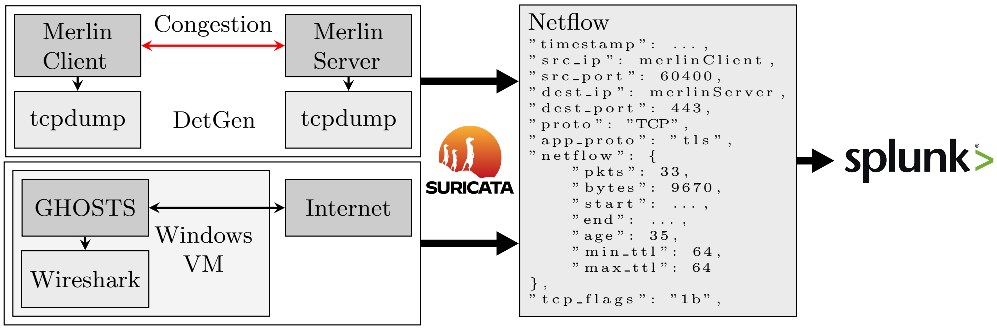
[11] https://github.com/BishopFox/sliver/

[12] https://github.com/redcanaryco/atomic-red-team

[13] https://github.com/Ne0nd0g/merlin

[14] https://docs.docker.com/compose/

the network traffic using `tcpdump`. This separation into container is then utilised when labelling traffic. In the experiment, the Merlin client and server ran in separate containers, with their associated ”`tcpdump`-containers” connected directly to the network interface of the Merlin containers In addition, the DetGen framework adds congestion and other small errors to make the simulation more realistic. This is illustrated in figure 1 (top-left).



```
"timestamp": ...,
"src_ip": merlinClient,
"src_port": 60400,
"dest_ip": merlinServer,
"dest_port": 443,
"proto": "TCP",
"app_proto": "tls",
"netflow": {
    "pkts": 33,
    "bytes": 9670,
    "start": ...,
    "end": ...,
    "age": 35,
    "min_ttl": 64,
    "max_ttl": 64
},
"tcp_flags": "1b",
```

Fig. 1. Experimental setup of Merlin with DetGen, and GHOSTS

The setup is sufficient to train a ”signature-model” which can recognise C&C-traffic. However, if our aim is to use supervised learning to train a classifier, we also need benign traffic. To achieve this, we used a framework called General Hosts (GHOSTS) [8] to simulate benign traffic. We configured GHOSTS to visit a list of domains known to be benign and record the traffic (using Wireshark), meaning, as with Merlin, we are capturing HTTP-traffic. Due to technical reasons and time constraints, we did not integrate GHOSTS into the DetGen framework, and instead captured GHOSTS traffic separately in a Windows virtual machine. This is sufficient for our proof-of-concept but will need to be integrated in the future. GHOSTS was then used to to connect to live domains with real world congestion applied. Figure 1 shows the full experimental setup for Merlin and GHOSTS. After running the simulation we changed the IP address of the Merlin Client to be the same as the GHOSTS Client. We then used Suricata[15] to convert the data sets to Netflow before combining them and importing them to Splunk[16] for visualisation. Up to this point, the C&C and Merlin traffic were in separate files, which we exploited when labelling during import into Splunk.

Splunk can then be used to train classification models, which can further be applied to real data. Here, we only visualise the traffic to illustrate how the labels can separate the traffic, as shown in figure 2. The plot shows both the number of bytes transmitted from the Merlin Client to the Merlin Server and benign traffic generated by GHOSTS (in logarithmic scale). It is easy to see the heartbeat in the graph from the Merlin Client. Note that there is some discrepancies initially

[15] https://suricata.io/

[16] https://www.splunk.com/

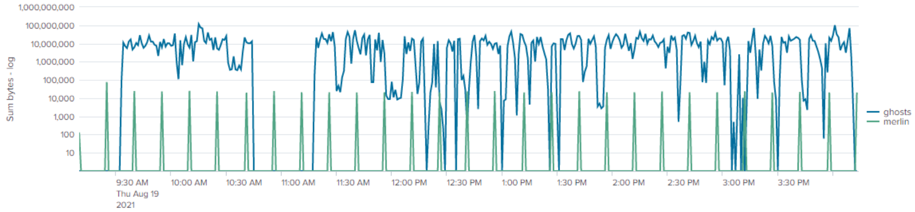


Fig. 2. #bytes of Merlin C&C traffic compared to traffic from GHOSTS in Splunk.

in the heartbeat due to some issues during setup, while a sudden dip in GHOSTS is likely caused by a need for user input to a CAPTCHA or similar.

3 Discussion and further work

We have shown that C&C tools used in practice can be simulated and labelled in a way that it can be separated from benign traffic in a SIEM with a fine grain of atomicity, which can further be utilised to train machine learning models for NIDS. Whilst the scientific contribution presented here may be limited, we believe our approach is promising for applying underlying research in an operational setting. This will require that the simulations are ran over longer time periods, using different C&C tools, different configuration and different architectures. This is also the case for the simulated benign traffic, where GHOSTS need to be integrated into the DetGen framework.

References

1. Buchanan, M., Collyer, J.W., Davidson, J.W., Dey, S., Gardner, M., Hiser, J.D., Lang, J., Nottingham, A., Oprea, A.: On generating and labeling network traffic with realistic, self-propagating malware. arXiv preprint arXiv:2104.10034 (2021)
2. Clausen, H., Flood, R., Aspinall, D.: Traffic generation using containerization for machine learning. arXiv preprint arXiv:2011.06350 (2020)
3. Garcia, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. computers & security **45**, 100–123 (2014)
4. Hutchins, E.M., Cloppert, M.J., Amin, R.M., et al.: Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research **1**(1), 80 (2011)
5. Landauer, M., Skopik, F., Wurzenberger, M., Hotwagner, W., Rauber, A.: Have it your way: Generating customized log datasets with a model-driven simulation testbed. IEEE Transactions on Reliability **70**(1), 402–415 (2020)
6. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: Rapid training data creation with weak supervision. In: Proc. of the VLDB Endowment. vol. 11, p. 269. NIH Public Access (2017)
7. Recorded Future by Insikt Group: Adversary infrastructure report 2020: A defender's view. Tech. Rep. CTA-2021-0107 (2021)
8. Updyke, D., Dobson, G., Podnar, T., Earl, B., Cerini, A.: Ghosts in the machine: A framework for cyber-warfare exercise npc simulation. Tech. Rep. CMU/SEI-2018-TR-005, SEI/CMU (December 2018)

Comparing Natural and Strong Typing Behavior for Keystroke Dynamics Multimodal Database Collection*

Markus Hoff Skudal[1], Matus Pleva[2][0000−0003−4380−0801], Štefan Korečko[2][0000−0003−3647−6855], Patrick Bours[1][0000−0001−5562−6957], and Daniel Hladek[2][0000−0003−1148−3194]

[1] Norwegian University of Science and Technology,
markuhs@stud.ntnu.no; patrick.bours@ntnu.no
[2] Technical University Kosice;
{Matus.Pleva;Stefan.Korecko;daniel.hladek}@tuke.sk

Abstract. Multimodal Keystroke Dynamics has been accurately shown to increase identification and authentication precision. When expanding this technique with EMG (Electromyography) analysis, a concern arose regarding a suspected low dynamic range of the EMG data. Thus a small-scale, multimodal experiment with five subjects was performed to display its potential, both with natural and strong typing behavior. This experiment lays the groundwork for a future multimodal Keystroke Dynamics project with 50–100 subjects. The results suggest that natural typing behavior gives high quality, and possibly better, data than strong typing. For 1-session-training, the evaluation shows user identification accuracy of 93% for natural and 85% for strong typing behavior. Hence, further research of scale could rely on natural typing behavior as the standard approach of recording.

Keywords: Keystroke Dynamics · User Identification · EMG · Timing Analysis · Typing Behavior.

1 Introduction

The experiment presented in this poster captured multimodal data in the form of (1) Keyboard timing, (2) EMG signals from Myo[3] armbands, (3) audio, and (4) video. More details on timing and audio keyboard analysis background can

* This work has been supported by the Slovak Research and Development Agency projects APVV-16-0202 & APVV-15-0517, by Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under the research projects VEGA 1/0753/20 & VEGA 2/0165/21 and by Cultural and Educational Grant Agency of the Slovak Republic grant Nos. KEGA 009TUKE-4/2019 & KEGA 035TUKE-4/2019, funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic.

[3] https://developerblog.myo.com/

be found in [4]. From this data, timing and EMG data sets were studied to determine typing behavior effectiveness. Recording of the Keystroke Dynamics data was exercised in four sessions for each of the five subjects, for each typing behavior as described in [3]. This resulted in 20 session recordings for each form of typing. On average these sessions lasted 80 seconds and recorded 25 correct typings of a predetermined word: ”password”. Ideally, the experiment should have had a larger subject base. This, however, proved difficult due to the COVID-19 pandemic and local, physical restrictions.

1.1 Keyboard Timing Analysis

In Keyboard Timing Analysis we measure when each key is pressed down and when it is released. From that, we can calculate the time a key is held down (called the duration of that key) and the time elapsed between typing two consecutive keys (called latency between the keys). Because the main goal of this preliminary research was to test the capability to use EMG for identification, we have only applied simple statistical analysis of the timing data, similar to what has been done by Pleva et al. in [4]. Besides the Scaled Manhattan Distance (SMD) that was used in [4], we also implemented the Scaled Euclidean Distance (SED), as these are rather similar in their implementations.

1.2 EMG Data Analysis

For analysis, the EMG data was loaded from CSV files into a data-handler for easy access and manipulation. With one Myo armband on each arm, and each armband producing eight EMG signals, each timestamp consisted of 16 data points. The sampling rate was found to vary between 165–175 Hz, though it originally was set to 200Hz. In order to compare the two forms of typing behavior, we chose MFCC (Mel-frequency cepstrum coefficients) [4, 5] with a 2-second window frame and 0.5-second step size to retrieve features. The MFCC was performed on each signal individually, with a dynamically adjusted sampling rate, and then the coefficients were combined in a 1x208 dimension array for each time frame (16 signals * 13 coefficients).In total, the experiment resulted in 2806 samples with 208 features, which stem from approximately 140 samples from each subject per session.

Despite our modest sample size, we hypothesized that the feature extraction from the MFCC would provide pattern rich samples for neural network analysis. We experimented with four Keras[4] models: GRU (Gated Recurrent Units), LSTM (Long Short-Term Memory), CNN 1D (one-dimensional Convolution Network), and a basic Feed-Forward-Network for reference. The hypothesis was demonstrated as plausible, as all networks cross-validated performed above 85%. The highest performing one-dimensional CNN-model was further used to compare natural and strong typing behavior. This model is composed of a Conv1D (32 nodes) and MaxPooling1D (5-sized kernel) layer, together with

[4] https://keras.io/

a Flatten, Dense (128 nodes), and Dense (5 nodes/classes, Softmax) layer. The following results are based on parameters batch size and epochs set to 64 and 30 respectively.

2 Results

2.1 Keyboard Timing Results

As mentioned above, we have used the SMD and SED distance metrics to evaluate the performance of the timing information. We have tested it on duration values only, latency values only, as well as the full set of duration and latency values together. The results are given in Table 1.

Table 1. Performance results from KD timing information

Typing Behavior:	**Natural**		**Strong**	
	SMD	SED	SMD	SED
Duration only	0.77	0.77	0.49	0.66
Latency only	0.55	0.63	0.74	0.76
All features	0.73	0.78	0.65	0.78

The results in Table 1, even though only based on the typing behavior of 5 persons, seem to indicate that duration values give a better performance when using natural typing, while latency features perform better for the strong typing behavior. When using the combination of duration and latency features we see that the natural typing behavior has a better performance when using SMD while for SED the performance is the same for both types of typing behavior.

Based on the results of the typing behavior there is no clear advantage in collecting strong typing behavior samples in addition to the normal typing behavior samples.

2.2 EMG Analysis Results

EMG Data Characteristics. From the raw EMG data, we retrieved important characteristics to describe its properties and compared the data from natural and strong typing behavior. Maximum and Minimum values were studied, together with the Median, Mean Absolute ($MA = \frac{1}{n}\sum_{t=1}^{n}|x_t|$), and Root Mean Squared ($RMS = \sqrt{\frac{1}{n}\sum_{t=1}^{n}x_t^2}$) [1, 2]. It shows that on average both were higher for *strong* typing behavior: MA 6.9 vs *7.4* and RMS 11.6 vs *12.4*.

Identification performance. When comparing the ability for identification using the best performed CNN 1D, both data sets provided above 85%. When training on one session the Natural typing behavior scored a good 92.8% while the Strong typing behavior gave an accuracy of 85.2% (see Fig. 1).

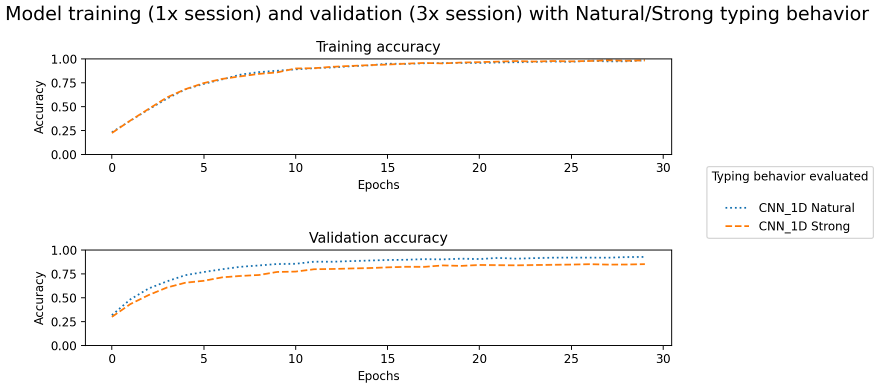


Fig. 1. The plot is the result of cross-validated session training on each type of typing behavior. The average of identification accuracy for validation gave a 92.8% vs. 85.2% with 1-session-training.

3 Discussion & Conclusion

As presented in this paper, the results suggest no clear benefit for further usage of Strong typing behavior in Multimodal Keystroke Dynamics Analysis. The indications, however, are not entirely aligned in this view. While EMG sensory identification shows beneficial results for the exclusive use of Natural typing behavior, the timing analysis is ambiguous. Based on these findings, we decided to efficiently allocate resources to Natural typing behavior only. Hence, further research with Multimodal Keystrokes Dynamics on 50–100 subjects will utilize this methodology in its recording.

References

1. Chong, H.J., Kim, S.J., Yoo, G.E.: Differential effects of type of keyboard playing task and tempo on surface EMG amplitudes of forearm muscles. Frontiers in Psychology **6**, 1277 (2015). https://doi.org/10.3389/fpsyg.2015.01277
2. Grabczyński, A., Szklanny, K., Wrzeciono, P.: Applying EMG technology in medial and lateral elbow enthesopathy treatment using Myo motion controller. Australas Phys Eng Sci Med. **42**(3), 701–710 (2019), doi.org/10.1007/s13246-019-00770-5
3. Korečko, Š., Pleva, M., Haluška, M., Skudal, M.H., Bours, P.: EMG input data collection for multimodal keystroke analysis. In: 12th IEEE Int. Conference on Cognitive Infocommunications, CogInfoCom 2021 Sept. 23-25. pp. 205–210. IEEE (2021)
4. Pleva, M., Bours, P., Ondáš, S., Juhár, J.: Improving static audio keystroke analysis by score fusion of acoustic and timing data. Multimedia Tools and Applications **76**(24), 25749–25766 (2017). https://doi.org/10.1007/s11042-017-4571-7
5. Rahim, M.A., Shin, J.: Hand movement activity-based character input system on a virtual keyboard. Electronics **9**(5), 774 (May 2020). https://doi.org/10.3390/electronics9050774