Proceedings of the 4th Seminar on

# Research Trends in Media Informatics

Institute of Media Informatics
Ulm University

EDITORS
Naim Asaj, Bastian Könings, Mark Poguntke, Florian Schaub, Björn Wiedersheim and Michael Weber

14th February 2012

CONTENTS

# Foreword

RTMI'12 – the fourth incarnation of the seminar *Research Trends in Media Informatics* – offers a program full of exciting papers covering important research topics in Media Informatics and related fields. At this year's seminars the papers concentrate on three main topics: security and privacy, HCI, and novel networking technologies and paradigms.

Mobile Internet access and wireless communications see an increasing adoption, as do cloud services and cloud storage. Security research is essential to ensure that mobile clients can securely communicate with the cloud. The papers in the security and privacy session highlight new protocols for authenticating clients and authorizing data access, discuss security issues in our mobile communication systems, and address the security and privacy challenges posed by the future Internet of the Things.

The HCI session reflects the diversity of HCI research quite well. Building user interfaces that convey important information but do not distract drivers is key for automotive HMI systems. The first paper gives an overview of the requirements that driver distraction avoidance poses on car interfaces and how they can be realized. As a contrast, the second paper looks at the emerging trend of gamification. The integration of game elements in the user experience can lead to stronger and longer acting engaged of users with a system or service.

The trend towards miniaturization of devices and the increasing number of gadgets we carry around with us introduces the potential of distributed and self-organizing sensor networks with a wide variety of applications. The papers in the networking technologies session discuss two types of such networks that hold great potential. Body area networks connect devices and sensors on the human body to enable health monitoring with higher precision. Flying ad-hoc networks connect swarms of self-organizing micro air vehicles (MAVs) for different purposes. Both network types poses unique challenges on communication networks and protocols which are discussed by the respective papers.

The RTMI seminar series aims to motivate students to delve deeper into the vast and diverse research in the Media Informatics domain, going beyond topics usually covered in lectures. Participants chose one of many topics that were pre-selected to highlight directions and challenges, which are focus of active research – not only at the Institute of Media Informatics at Ulm University, but around the world. In order to provide participants with insights into the academic publishing cycle, the seminar emulates the process of submitting a technical paper to an academic conference. Participants first prepared their papers on a selected topic, either in English or German. The submissions were then peer reviewed in terms of content, academic quality, and presentation by at least two other seminar participants and one editor. The revised camera ready versions of the contributions constitute these proceedings. The authors presented their papers at the RTMI conference in February 2012. The RTMI conference was held in the remarkable Hall of Knights at the Villa Eberhardt in Ulm, which provided a perfect setting for high quality presentations, interesting discussions, and debates.

The editors would like to thank all authors for their effort and the work put into each individual contribution.

Ulm, February 2012

*Florian Schaub, Naim Asaj, Bastian Könings,*
*Mark Poguntke, Björn Wiedersheim, and Michael Weber*

# Conference Program

Tuesday, February 14, 2012

9:00 *Welcome*
9:15 **Keynote**

10:15 *Coffee Break*

**Session I: Security and Privacy**
*Chair: Bastian Könings*

10:30 Authentifizierung und Autorisierung mit OpenID und OAuth ~ *Daniel Ritter*
11:00 GSM/3G/4G/DECT Security ~ *Dominik Deuter*
11:30 RFID Security and Privacy ~ *Philipp Hock*

12:00 *Lunch Break*

**Session II: Human Computer Interaction**
*Chair: Naim Asaj*

13:30 Driver Distraction in Automotive HMI ~ *Marcel Reichersdörfer*
14:00 Gamification: State of the Art Definition and Utilization ~ *Fabian Groh*

14:30 *Coffee Break*

**Session III: Networking Technologies**
*Chair: Björn Wiedersheim*

14:45 Body Area Networks ~ *Jonathan Sondershaus*
15:15 Flying Ad-Hoc Networks ~ *Michael Müller*

15:45 *Coffee Break*

16:00 Closing Remarks & Best Paper Award

18:00 *Social Event*

# Authentifizierung und Autorisierung mit OpenID und OAuth

Daniel Ritter
Institute of Media Informatics
Ulm University
Ulm, Germany
Email: daniel.ritter@uni-ulm.de

***Abstract*—Dieser Bericht stellt im ersten Teil das Protokoll OpenID [1] und im zweiten Teil das OAuth-Protokoll [2] vor. Zunächst wird jeweils beschrieben für was und von wem die Protokolle eingesetzt werden. Wichtige Begriffe wie Token, Ressources etc., die zum Verständnis notwendig sind, werden im jeweiligen Kontext geklärt. Anschließend wird jeweils detailliert auf den Protokollfluss und die dabei ausgetauschten Daten eingegangen. Nach Kenntnis des Protokollflusses werden sicherheitstechnische Gesichtspunkte betrachtet. Im OpenID Teil wird generell auf Angriffspunkte und mögliche Gegenmaßnahmen eingegangen, wohingegen im OAuth Teil explizit eine bekannte Schwachstelle einer älteren OAuth Version beschrieben wird.**

## I. Einleitung

Unter den Top Ten der beliebtesten Webseiten [3] bietet ausnahmslos jede Webseite eine Anmeldung an. Wie soll man bei der Vielfalt an Webseitenaccounts noch den Überblick behalten? Eine mögliche Lösung für dieses Problem sind Single-Sign-On Lösungen. Hierbei wir ein Login für alle Webseiten genutzt.

Auf diesem Gebiet existieren bereits einige Lösungen, wie beispielsweise das von Microsoft 1999 eingeführte ".NET Passport" [4]. Einmal mit ".NET Passport" angemeldet, hatte man Zugriff auf bis zu 100 Partnerseiten; aber meistens nur Microsoft Dienste, wie Hotmail, Messenger etc. Seit 2007 wird ".NET Passport" unter dem Namen "Windows Live ID" [4] weiterentwickelt und kann ebenfalls von Dritten in eigene Lösungen integriert werden. Allerdings wurde bei diesem System die zentrale Speicherung der Profildaten in Verbindung mit der Monopol-Stellung von Microsoft kritisiert, was zu einer geringen Akzeptanz führte (sowohl Nutzer als auch die teilnehmende Internetseite müssen sich bei Microsoft registrieren).

Der nächste ernst zu nehmende Vorschlag zur Lösung des Kennwort-Dilemmas kam 2005 von Brad Fitzpatrick. Er stellte 2005 die Single-Sign-On Lösung OpenID der Öffentlichkeit vor. OpenID ist offen, dezentral und für Jedermann einsetzbar. Mittlerweile ist OpenID soweit im Netz verbreitet, dass es das Potential haben könnte zur ersten umfassenden Single-Sign-On Lösung des Internets zu werden. Versierte Internetnutzer haben davon wahrscheinlich bereits gehört oder nutzen es sogar bereits. Um Interessierten mehr Einblick in das Protokoll zu geben, wird sich der erste Teil des Papers mit OpenID auseinandersetzen. Im Kontext mit OpenID fällt auch oft der Begriff OAuth. Betrachtet man OpenID und OAuth nur oberflächlich, scheinen die Prozesse sehr ähnlich zu sein. Die Ähnlichkeit kommt jedoch nur dadurch zustande, dass mehrere Personen an der Entwicklung von beiden Protokollen beteiligt waren. OAuth ist ein von OpenID vollkommen unabhängiges Protokoll und wird sogar für einen vollkommen anderen Zweck eingesetzt. OpenID wird zur Authentifizierung und OAuth zur Authorisierung verwendet.

Auch wenn sich Authentifizierung und Autorisierung ähnlich anhören, muss man genau differenzieren, was damit im sicherheitstechnischen Bereich gemeint ist. Authentifizierung bedeutet, ob jemand wirklich die Person ist, für die er sich ausgibt. Mit Autorisierung ist hingegen gemeint, ob eine Person, die ggf. davor schon ihre Identität bewiesen hat, das Recht hat, auf eine bestimmte Entität zuzugreifen. Eine typische Fragestellung im OpenID-Kontext lautet beispielsweise: "Ist die Person, die sich einloggen will, tatsächlich Jane?". Die typische Frage im OAuth-Kontext lautet hingegen: "Darf der Fotodruckdienstleister Müller.de auf Janes flickr Fotos zugreifen?". Wie diese zwei Fragen aus technischer Sicht sicher beantwortet werden können, wird in den folgenden Kapiteln erläutert.

## II. Das OpenID-Protokoll

### A. Was ist OpenID?

OpenID ist ein offenes, dezentralisiertes Single-Sign-On-Protokoll zur Authentifizierung von Benutzern gegenüber Webseiten. Die Bezeichnungen *OpenID-Provider* und *Relying Party* sind zum Verstehen des OpenID-Protokolls notwendig. Eine Relying Party vertraut auf die Identitätsaussage eines OpenID-Providers. Für den Identitätsbeweis ist ein so genannter Identity Provider (IdP/Server/Provider) notwendig. Dies ist ein OpenID Authentication Server, der einen Beweis liefert, dass dem Benutzer die behauptete Identität gehört. Dezentralisiert heißt, dass keine zentrale Autorität nötig ist, die Relying Parties oder OpenID Provider prüft bzw. registriert.

Anstatt Benutzernamen nutzt OpenID einen URI als Identifier. Der Identifier ist entweder eine Subdomain des OpenID-Providers (z.B. "benutzername.myopenid.com") oder ein Pfad in der Provider URL (z.B. "myopenid.com/benutzername"). Soll der Identifier providerunabhängig sein, so kann man einen selbstdefinierten URI verwenden. Die Authentifizierung erfolgt dadurch, dass der Benutzer die Kontrolle über einen Identifier durch Eingabe eines korrekten Passwortes bestätigt.

Dies geschieht ohne dass die Relying-Party Zugriff auf vertrauliche Informationen, wie Passwort oder Email-Adresse (außer durch explizites Anfragen; zum Beispiel mittels SReg) erhält. OpenID beruht nur auf Standard HTTP(S)-Anfragen bzw. Antworten und setzt keine besonderen Fähigkeiten an moderne Browser voraus. Auch Javascript ist nicht notwendig für die OpenID-Fähigkeit. Jedoch arbeitet der Protokollfluss von OpenID sehr gut mit AJAX-Webseiten zusammen, so dass der Benutzer die momentane Webseite des Consumers nicht verlassen muss. Außerdem ist das Protokoll nicht auf Cookies oder andere Session-Management Unterstützungen angewiesen.

Da OpenID Open Source ist, kann jeder seinen eigenen OpenID-Server betreiben und als OpenID-Provider fungieren. Viele große Unternehmen wie Google, IBM, Microsoft, Myspace, PayPal, VeriSign oder Yahoo unterstützen den Standard und vergeben an neuregistrierte Benutzer automatisch entsprechende OpenIDs. Meldet man sich mit seiner URL-Identität und Passwort bei seinem OpenID-Provider an, so kann man sich theoretisch bei allen Webseiten, die OpenID unterstützen authentifizieren ohne seinen Benutzernamen und Passwort preisgeben zu müssen. In der Praxis sperren manche Provider allerdings den Zugang zu ihren Services durch fremde OpenID-Provider. In vielen Fällen ist das sinnvoll, so dass anonyme beziehungsweise Wegwerf-Provider nicht zu einem Login fähig sind[5].

Das OpenID-Konstrukt baut sich aus fünf Protokollen auf (siehe Tabelle I). Dabei ist das OpenID Authentication 2.0 Protokoll der Hauptbestandteil der Protokollfamilie und spezifiziert die Basis für die Authentifizierung. Die weiteren Protokolle in Tabelle I bauen auf diesem Grundstein auf und stellen Erweiterungen für anderen Datenaustausch dar. Das Protokoll *OpenID Simple Registration Extension 1.0 (SReg)* kann beispielsweise genutzt werden, um acht fest definierte Profilinformationen vom Provider zur Relying-Party zu übertragen. Ein Konsument und ein Provider dürfen keine Attribute ohne Zusage des Endbenutzers austauschen. Das gehört jedoch nicht zu den OpenID-Spezifikationen und ist beim Implementieren bzw. der Auswahl des Providers genau zu beachten. Im nächsten Abschnitt wird näher auf das *OpenID Authentication 2.0 - Final* zur Benutzeridentifikation eingegangen.

### *B. Protocol Workflow*

Zunächst wird genauer auf das Nachrichtenformat eingegangen, das bei der Kommunikation verwendet wird. Die Anfrage- und Antwortparameter werden in Key-Value-Form kodiert. Der Key enthält den Präfix “openid.” um Namensraumkonßikten durch weitere Extensions, die die gleichen Parameter verwenden könnten, vorzubeugen. Der Key ist vom Value durch einen Doppelpunkt getrennt. Nachrichten können über GET oder POST HTTP-Nachrichten übertragen werden. Wird eine valide Antwort empfangen, antwortet der Server mit dem Statuscode 200. Bei GET-Anfragen werden die Parameter in der URL als Query nach RFC 3986 [7] kodiert. Wird eine Nachricht per POST übermittelt, so werden die OpenID Paramter im POST Body im “x-www-urlencoded”-Format übertragen. Ein POST-Body könnte beispielsweise wie folgt aussehen: “openid.mode=error&openid.error=This%20is%20an-%20example%20message”. Alle Nachrichten müssen zwingend die zwei Keys *openid.ns* und *openid.mode* enthalten. Openid.ns spezifiziert die Protokoll-Version und ist bei OpenID Authentication 2.0 standardmäßig der Wert “http://specs.openid.net/auth/2.0”. Fehlt dieser Key komplett oder dessen Wert ist “http://openid.net/signon/1.1” bzw. “http://openid.net/signon/1.0”, soll die Nachricht im *OpenID Authentication 1.1 Compatibility mode* verarbeitet werden. Das gewährleistet eine Abwärtskompatibilität zu älteren OpenID Versionen.

Bei OpenID wird unterschieden zwischen direkter und indirekter Kommunikation. Mit direkter Kommunikation ist ein Nachrichtenaustausch zwischen Relying Party und Provider ohne Mittelsmann gemeint. Die direkte Kommunikation wird bei der Authentifizierung lediglich an zwei Stellen verwendet: zur Erstellung von Assoziationen (siehe Schritt 3) und zur Prüfung von Authentifizierungsantworten (siehe Schritt 7). Bei der indirekten Kommunikation werden die Nachrichten über den User-Agent zwischen Provider und Consumer vermittelt. Die indirekte Kommunikation kann sowohl vom Provider als auch vom Consumer gestartet werden. Die indirekte Kommunikation wird für Authentication-Anfragen und Authentication-Antworten verwendet. Technisch wird der indirekte Datenaustausch über HTTP Weiterleitungen oder das Absenden von HTML Forms realisiert. Die typische Authentifizierung eines Consumers besteht aus 7 Phasen, die in Abbildung 1 anschaulich illustriert sind und nachfolgend näher beschrieben werden.

*1) Starten der Authentifizierung:* Der Benutzer besucht im ersten Schritt eine Consumer-Webanwendung und wird dort aufgefordert seinen Identifier einzugeben. Der Benutzer hat in diesem Schritt die Möglichkeit zwei verschiedene Arten von Identifiern einzugeben – den OP Identifier oder den Claimed Identifier. Der OP Identifier ist ein Identifier für den OpenID Provider - also zum Beispiel *myopenid.com*. Wenn ein OP Identifier vom Endbenutzer angegeben wird, muss ein konkreter Claimed Identifier beim Provider ausgewählt werden, der für die Relying Party vewendet werden soll. Es ist also möglich mehrere Identifier beim gleichen Provider zu registrieren und bei Bedarf beim Provider mit einem Identifier-Auswahldialog konfrontiert zu werden. Dadurch können verschiedene Identifier für unterschiedliche Webseiten benutzt werden, die alle einheitlich bei einem Provider verwaltet werden können. Der Claimed Identifier ist ein Identifier über den der Benutzer behauptet die Kontrolle zu besitzen - beispielsweise “benutzername.myopenid.com”. Das gesamte OpendID-Authentication Protokoll ist dahingehend ausgelegt diese Behauptung zu verifizieren.

*2) Ermittlung des Providers:* Der Identifier wird anschließend normalisiert. Unter Normalisation ist zu verstehen, dass zum Beispiel “http://” ergänzt wird, falls das Schema vom User nicht eingetippt wurde. Die Auffindung des korrekten Providers anhand des Identifiers nennt sich Discovery und

Table I
DER OPENID-PROTOKOLLSTACK BESTEHT AUS FÜNF SPEZIFIKATIONEN. ALLE FÜNF PROTOKOLLE ZUSAMMEN BILDEN DAS SOG. OPENID-FRAMEWORK. QUELLE: [6]

| Protokoll | Beschreibung |
|---|---|
| OpenID Authentication 1.1 Draft 1 und 2.0 | Authentifizierungsspezifikation, Dies ist der Hauptteil im OpenID-Protokoll |
| OpenID Simple Registration Extension 1.0 und 1.1 Draft 1 (SReg) | Austausch von Identitätsattributen mit einer festen Anzahl von acht vorgegebenen Attributen |
| OpenID Attribute Exchange 1.0 (AX) | Austausch von Identitätsattributen, unbegrenzt und selbst definierbar, erweiterbar, zukünftig wichtiger als SReg |
| OpenID Provider Authentication Policy Extension 1.0 (PAPE) | Überprüfung bestimmter Verifizierungsmethoden: Wurden ausschließlich Benutzername und Passwort oder zusätzlich Token-Generator verwendet? Hinzu kommt die Überprüfung der Verifizierungszeit des Providers bei dem angegebenen Benutzer. Unter verify.sxip.com kann man eine Demo-Seite für PAPE aufrufen |
| OpenID Assertion Quality Extension 1.0 Draft 3 | Überprüfung der Attributsqualität: Wurde die angegebene E-Mail-Adresse tatsächlich auf ihre Richtigkeit überprüft? Gehört die E-Mail wirklich der angegebenen Person? |

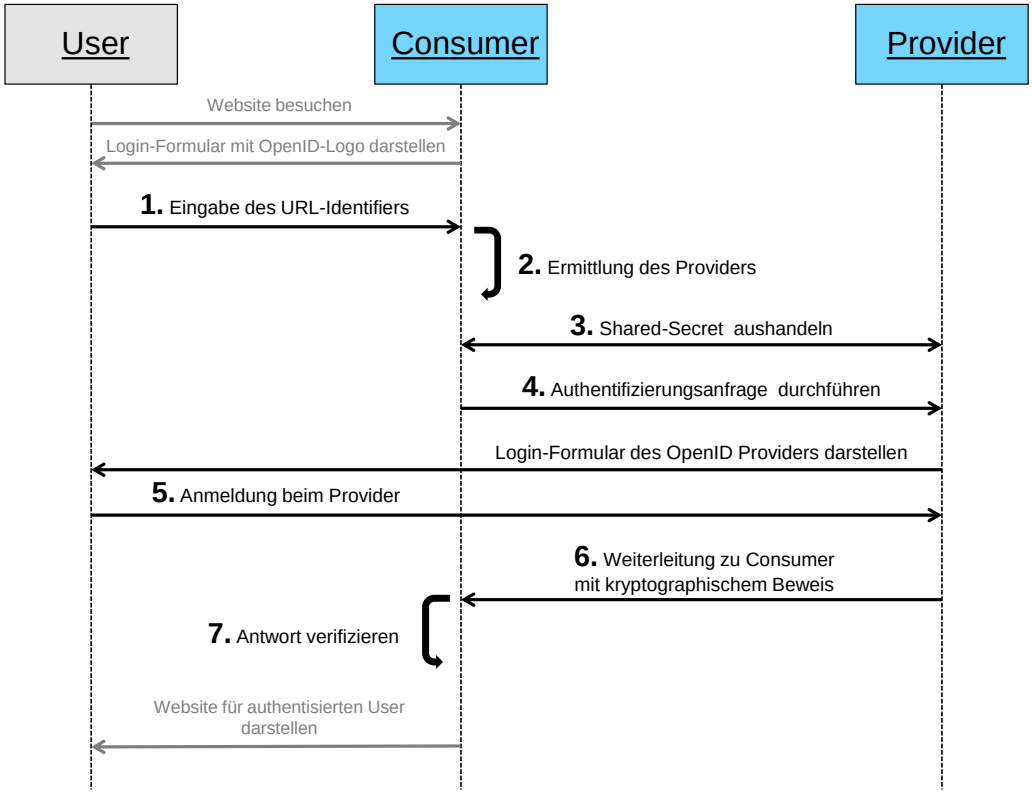


Figure 1. Protokollablauf einer Authentifizierung mittels OpenID. Zu den Nummern an den Pfeilen gibt es entsprechende Abschnitte im Text, die die Kommunikation genauer beschreiben.

liefert ein XML-Dokument[1] mit der sogenannten *OpenID Endpoint URL* und anderen Informationen zurück. Alle nachfolgenden Authentication Anfragen des Consumers werden an diese Endpoint-URL geschickt. Zur Auffindung der Endpoint URL spezifiziert OpenID drei Verschiedene Möglichkeiten: XRI Resolution 2.0 [8], Yadis [9] und HTML-Based Discovery. XRI Resolution wird eingesetzt, falls es sich bei dem Identifier um eine XRI handelt. Hat der Identifier die Form einer URL wird das Yadis Protokoll gestartet. Beide Verfahren liefern ein XRDS-Dokument zurück. Wenn das Yadis Protokoll fehlschlägt, wird versucht den Identifer mittels HTML-Based Discovery aufzulösen. Bei der HTML basierten Auffindung wird versucht den Identifier als URL zu einer Webseite zu interpretieren. Gelingt der Abruf der Webseite, so wird deren Head-Bereich nach bestimmten Meta-Tags geparst, um Informationen wie die Endpoint URL und andere Daten zu ermitteln. Zu den ermittelten Informationen gehört neben der OP Endpoint URL auch immer die *Protocol Version*. Wenn der Benutzer in Schritt 1 einen Claimed Identifier eingegeben hat, so stehen auch die Daten Claimed Identifier und *OP-Local Identifier* zur Verfügung. Der OP-Local Identifier wird vom Provider als alternative, interne Repräsentation des Claimed Identifier eines Benutzers verwendet. Hat der Benutzer jedoch einen OP Identifier eingegeben steht zu diesem Zeitpunkt kein Claimed Identifier bereit. Für nachfolgende Authentifizierungsanfragen an den Provider muss deshalb der Claimed Identifer und der OP-Local Identifier auf "http://specs.openid.net/auth/2.0/identifier_select" gesetzt werden. Dadurch wird dem End-User in Schritt 5 ein Auswahldialog seiner registrierten Identifier angezeigt.

*3) Shared-Secret zwischen Relying Party und Provider aushandeln:* Der dritte Schritt ist optional und ist notwendig für die *indirekte Verifikation* von Antwortnachrichten. In diesem Schritt wird eine Assoziation zwischen Konsument und Provider in Form eines Shared-Secrets hergestellt. Der Assoziationsvorgang wird durch eine direkte Anfrage des Consumers an den Provider mit dem Typ *openid.mode:associate* gestartet. Das Shared-Secret wird durch das Diffie-Hellman Key Exchange Verfahren [10] auf beiden Seiten erzeugt. Der Provider nutzt den Geheimschlüssel, um nachfolgende Nachrichten zu signieren. Die Relying Party kann dadurch die Integrität der Nachrichten überprüfen. Somit kann ein Zustand beibehalten werden, so dass bei späterer Kommunikation kein weiterer direkter Kontakt zwischen Relying Party und Consumer notwendig ist, um die Integrität der Antwortnachrichten zu überprüfen. Wird dieser Schritt übersprungen, so muss jede Antwortnachricht des Providers durch eine Verifikationsanfrage an den Provider überprüft werden, wodurch sich die Roundtriptime entsprechend verlängert. Dies nennt sich *direkte Verifikation* und wird mit einer Anfragenachricht des Typs *openid.mode:check_authentication* und einer Kopie aller zuvor Empfangenen Key-Value-Paare (außer openid.mode) durchgeführt.

[1]Beim XML-Dokument handelt es sich konkret um ein XRDS-Dokument (Kurzform für eXtensible Resource Descriptor Sequence). XRDS ist ein simples generisches XML-Format um Web Ressourcen zu beschreiben und abrufbar zu machen.

*4) Weiterleitung zum Provider:* Hat die Relying Party den Discovery Prozess erfolgreich beendet und optional eine Assoziation hergestellt, so kann nun eine Authentication-Anfrage an den Provider geschickt werden. Die Authentication-Anfrage ist, wie bereits erwähnt, eine indirekte Anfrage d.h. der Consumer leitet den End-User mit einer Authentication-Request-Nachricht an den Provider weiter. Für die Anfragenachricht sind im Feld *openid.mode* die Werte "checkid_immediate" oder "checkid_setup" erlaubt. Wenn der Consumer eine Interaktion zwischen Relying Party und End-User erwünscht, so hat das Feld den letzteren Wert. Ist hingegen keine Interaktion gewünscht, wie dies zum Beispiel bei einer asynchronen Javascript Authentifizierungsanfrage der Fall ist, so hat das Feld den Wert "checkid_immediate".

*5) Anmeldung des Benutzers beim Provider:* Der Provider übernimmt die Authentifizierung des End-Users. In diesem Schritt kann der Benutzer unter Umständen (siehe Schritt 1) auswählen, welchen Identifier er für die Relying Party verwenden möchte. Welche Art der Authentisierung der Provider vornehmen muss, ist nicht im OpenID Framework spezifiziert.

*6) Weiterleitung zurück zur Relying Party:* Der Provider leitet den User-Agent zurück zum Consumer. Die Authentifizierungs-Antwort (auch genannt Assertion) sagt entweder aus, dass die Authentifizierung erfolgreich war (positive Assertion) oder enthält eine Fehlermeldung (negative Assertion). Das erfolgreiche Antwortformat enthält des Weiteren die Felder *openid.signed* und *openid.sig*, so dass der Client bei Durchführung von Schritt 3 die Integrität der Antwort prüfen kann. Das Erstere enthält eine Liste aller Keys, deren Werte in die Signatur eingegangen sind. Das letztere Feld enthält die vom Provider berechnete Signatur selbst. Als Signatur Verfahren stehen die Message Authentication Code (kurz MAC) Varianten HMAC-SHA1 oder HMAC-SHA256 zur Verfügung. Das konkrete Verfahren wird in der Authentication-Anfrage des Consumers durch das Feld *openid.assoc_type* ausgewählt. Ein MAC dient dazu die Integrität von Daten bzw. Nachrichten auf Seite des Consumers überprüfbar zu machen. Ein MAC-Algorithmus hat prinzipiell mindestens zwei Eingabeparamter: die *zu schützenden Daten* und einen Geheimschlüssel. Als Ergebnis liefert der MAC-Algorithmus eine Prüfsumme, die im Feld *openid.sig* als Wert übertragen wird. Die *zu schützenden Daten* werden in OpenID durch Konkatenieren aller Werte, deren Keys in openid.signed aufgelistet sind, erzeugt. Wurde in Schritt 3 kein Geheimschlüssel ausgehandelt, so verwendet der Provider einen zufälligen Wert als Geheimschlüssel, der gespeichert werden muss, so dass eine spätere direkte Verifikation möglich ist. Um Replay Attacken zu verhindern wird der Antwort eine Nonce beigefügt. Die Nonce muss mit dem aktuellen Zeitstempel des Servers beginnen und kann durch eine zufällige, beliebig lange Folge von Buchstaben erweitert werden. Ein Beispiel für eine korrekt formatierte Nonce wäre "2005-05-15T17:11:51ZUNIQUE".

*7) Verifizierung der Antwortnachricht:* Der Comsumer verifiziert die empfangene Authentifizierungsnachricht des Providers durch eine Überprüfung der Absender-URL, durch eine Verifizierung des erhaltenen Claimed Identifier, durch Überprüfung der Nonce und durch eine Überprüfung der

Signatur mit Hilfe des Geheimschlüssels (sofern Schritt 3 durchgeführt wurde). Wurde Schritt 3 nicht durchgeführt, so muss eine direkte Anfragenachricht an den Provider geschickt werden, um festzustellen, ob der Benutzer tatsächlich authentifiziert werden konnte. Anschließend ist der Benutzer beim Consumer eingeloggt.

Mit Überprüfung der Absender-URL ist folgendes gemeint: Bei einer Authentifizierungsanfrage (Schritt 4) wird der Key *openid.return_to* mit der URL der Relying Party als Value mitgeschickt. Für die Antwort wird dieser Wert eins zu eins kopiert und im gleichen Feld zurückgesendet. Nur wenn der empfangene Wert mit der URL, die der Consumer zuvor verschickt hat, übereinstimmt, wird die positive Assertion akzeptiert. Der erhaltene Claimed Identifier muss mit dem zuvor im Discovery Prozess ermittelten Claimed Identifier übereinstimmen. Ist der Claimed Identifier für den Consumer unbekannt, so muss mit dem erhaltenen Claimed Identifier der Schritt 2 erneut ausgeführt werden, um sicherzustellen, dass der Provider auch berechtigt ist, für diesen Identifier eine positive Assertion auszustellen.

### C. Sicherheitstechnische Gesichtspunkte

In diesem Abschnitt wird auf mögliche Schwachstellen von OpenID eingegangen und wie gegen diese vorgegangen werden kann.

*1) Eavesdropping Attacks:* Schneidet ein Angreifer die Kommunikation des OpenID Authentifizierungsprozesses mit, so stellt dies zunächst keinen ausnutzbaren Schwachpunkt im Protokoll dar. Wenn die Nonce allerdings nicht geprüft wird, so kann ein Angreifer die positive Authentifizierungsantwort abfangen und wiederverwenden. Diese Art von Angriff kann durch eine Verschlüsselung der Transportschicht verhindert werden.

*2) Man-in-the-Middle Attacks:* Durch das Signieren der Authentifizierungsantwort ist eine Verfälschung durch einen Man-in-the-Middle Angriff ausgeschlossen. Eine Änderung der signierten Felder ist ohne den geheimen Schlüssel nicht möglich, wenn das MAC Verfahren sicher ist. Für die eingesetzten MAC Verfahren sind bisher keine Schwachstellen bekannt. Bei MAC Verfahren ist zu beachten, dass die Geheimschlüssel möglichst zufällig gewählt werden, weil diese sonst durch eine gewisse Vorhersehbarkeit zu einer Schwachstelle führen könnten. Allerdings ist das sog. Tampering[2] in den Schritten 1, 2 und 7 möglich, da deren Felder nicht mit einem geheimen Schlüssel signiert sind. Wenn ein Angreifer das XRDS-Dokument beim Discovery-Prozess manipulieren kann (beispielsweise durch Kompromittieren der DNS Außösung oder der Transportschicht), so kann er sich dort selbst als Provider Endpoit ausgeben. Dadurch kann ein Passwort abgefangen werden ohne dass ein Man-in-the-Middle Aufbau nötig wäre. Eine Möglichkeit diese Art von Angriffen zu unterbinden ist es das XRDS-Dokument zu signieren [11]. Allerdings obliegt es der Relying Party den verwendeten Signaturen zu trauen oder nicht.

Eine andere Gegenmaßnahme ist es die DNS-Außösung mit einem SSL Zertifikat einer vertrauenswürdigen Zertifizierungsstelle (CA) abzusichern. Einen SSL Provider für einen Man-in-the-Middle Angriff zu imitieren, setzt voraus das Zertifikat zu stehlen oder zu fälschen, was bedeutend schwerer ist als netzwerkbasierte Angriffe. Für einen Sicherheitsgewinn muss SSL im gesamten Authentifizierungsablauf verwendet werden.

*3) Rogue Relying Party Proxying:* Mit Rogue Reyling Party Proxying ist nichts anderes gemeint als eine spezielle Phishing-Methode. Die böswillige Relying Party leitet den Anwender nach der Discovery nicht zum korrekten Provider weiter, sondern zu sich selbst, um das Passwort abzugreifen. Als Gegenmaßnahme könnte der Anwender ein eigenes Bild beim Provider hinterlegen, das bei der Login-Maske eingeblendet wird. Dadurch kann der Benutzer Phishingseiten entlarven. Unter Umständen können auch Browser Plugins eingesetzt werden, die die OpenID Server Endpoit URL verifizieren und bei einer Nichtübereinstimmung den Prozess abbrechen. Eine weitere Richtlinie der OpenID Spezifikation verlangt, dass Popups in einem neuen Fenster mit eingeblendeter Adressleiste geöffnet werden, so dass der Benutzer nicht getäuscht werden kann.

*4) Denial of Service Attacks:* Im OpenID Protokoll steht keine Möglichkeit bereit authentische von sinnlosen Anfragen zu unterscheiden, was zu einer Verwundbarkeit gegenüber DoS-Angriffen führt. Die anfälligste Phase ist die Aushandlung des Geheimschlüssels in Schritt 3, weil dies die meiste Last auf Seiten des Severs generiert. Diesem Problem kann mit Blocking-Maßnahmen, wie der IP-basierten Anfragelimitierung, entgegengewirkt werden.

## III. Das OAuth-Protokoll

### A. Was ist OAuth?

OAuth [2] ist ein Protokoll um fremde “Programme” (sog. Third-Parties) zu autorisieren auf zentral verwaltete Benutzerdaten, wie Fotos, Videos, Kontaktlisten etc. (sog. Ressourcen), zuzugreifen ohne die eigentlichen Authentifizierungsdaten (sog. Credentials) den fremden “Programmen” preisgeben zu müssen. Mit “Programmen” können auch andere Webservices gemeint sein. In anderen Worten heißt dies, dass OAuth ein Protokoll zur Zugriffsdelegation ist. Zum Beispiel kann Jane (Resource Owner) dem Müller-Druckdienst (Client/Consumer) Zugriff auf ihre bei ßickr gehosteten Bilder gewähren, ohne den Benutzernamen und das Passwort dem Müller-Druckdienst mitteilen zu müssen. Stattdessen authentifiziert sie sich direkt bei ßickr und ßickr erteilt dem Druckdienst die Zugriffserlaubnis auf ihre privaten Ressourcen. Daher der Begriff Zugriffsdelegation, da die Zugriffserlaubnis nicht direkt vom Benutzer an den Müller-Druckservice übertragen wird.

Für die weitere Beschreibung des OAuth-Protokolls müssen zunächst einige Begriffe, die teilweise im oberen Absatz bereits vorweg genommen wurden, geklärt werden: *Service Provider*, *Consumer*, *User/End-User*, *Protected Ressources*, *Token* und *Token-Secret*. Ein Service ist eine Webapplikation, die einem Consumer Zugriff auf Ressourcen mittels OAuth bereitstellt. Protected Ressources stellen Entitäten dar, für die Zugriffsrechte erteilt bzw. wiederrufen werden können. Anstatt einem Benutzernamen und einem Passwort werden für

[2]Nachrichtenverfälschung

die Identifizierung in OAuth so genannte Tokens eingesetzt. Tokens sind nichts anderes als base64-enkodierte Strings. Das Token-Secret ist stellvertretend für ein Passwort und das Token für einen Benutzernamen. Mit End-User oder User wird ein menschlicher Benutzer bezeichnet.

OAuth ist aber auch nützlich, um von verschiedenen Clients auf seine Ressourcen zugreifen zu können. Es existieren zwei Varianten des OAuth-Protokolls. Die dreibeinige OAuth-Variante, die aus Consumer, End-User und Provider besteht und damit eine Benutzerinteraktion erfordert. Daneben gibt es auch noch eine zweibeinige OAuth-Variante aus Consumer und Provider (also ohne Benutzerinterkation), die in der Praxis bereits auch schon häufig eingesetzt wird. In dieser Arbeit wird nur auf den dreibeinigen OAuth eingegangen, denn für die zweibeinige Variante gibt es noch keine offizielle Spezifikation und der zweibeinige OAuth ist nur ein Sonderfall des dreibeinigen OAuth.

Anstatt eines Benutzernamens und eines Passworts nutzt OAuth ein sogenanntes Token für die Autorisierung von Consumern. Ein Token ist ein eindeutiger Identifier, der vom Server/Provider an den Client/Consumer ausgestellt wird. Ein Token besteht meistens aus einem zufälligen String, so dass es eindeutig und schwer zu erraten ist. In den String ßießt zudem noch das ausgehandelte Shared-Secret mit ein, so dass Rückschlüsse zulässig sind, zu welchem Client das Token gehört. In OAuth existieren zwei unterschiedliche Token-Typen: Anfrage- und Zugriffs-Token. Mit dem Consumer werden also nicht die Credentials aus Benutzername und Passwort, sondern generierte Tokens geteilt. Das hat den Vorteil, dass für Tokens eine eingeschränkte Zugriffserlaubnis (d.h. man könnte einem Token nur Zugriff auf einen Teil der Ressourcen gewähren) und eine begrenzte Gültigkeit erteilt werden kann. Außerdem können Tokens unabhängig voneinander widerrufen werden, um die Autorisierung gezielt einzelnen Clients zu entziehen.

### *B. Protocol Workflow*

Bevor ein Consumer mit einem Provider kommunizieren kann, muss ein OAuth Entwickler noch einen zusätzlichen als separat ansehbaren Schritt erledigen. Der Entwickler muss sein "Programm" beim Provider registrieren und einen Consumer Key und ein Consumer Secret beantragen. Der Consumer Key identifiziert das Programm eindeutig und ist in jeder Anfrage anzugeben. Bei der Registrierung kann es sein, dass der Entwickler neben Name, URL und Autor noch einen Scope angeben muss. Mit einem Scope lässt sich der Zugriff auf bestimme Ressourcen bzw. Services einschränken bzw. limitieren. Dabei ist zu beachten, dass der Consumer Key und das Consumer Secret an den Scope gebunden sind. Sollte sich also der Scope eines Programmes später ändern, so ist eine Neuregistrierung erforderlich. Dieser Registrierungsprozess ist nicht Teil der OAuth Spezifikation und kann in der Service Provider Dokumentation nachgeschlagen werden. Für den weiteren Ablauf wird angenommen, dass der Entwickler diesen Schritt bereits erledigt hat.

Wie bereits erwähnt, nutzt OAuth zwei Typen von Tokens für unterschiedliche Zwecke – Access- und Anfrage-Token. Das Anfrage-Token wird verwendet, um die Autorisierungszustimmung des Benutzers zu repräsentieren. Mit einem Access-Token kann auf eine Resource zugegriffen werden. Die zwei Tokentypen haben auch ein korrespondierendes Geheimnis (Token-Secret) – es gibt ein Anfrage-Secret und ein Zugriffs-Secret. Es gibt also insgesamt drei Geheimnisse: Consumer Secret, Request Secret und Access Secret. Die Shared-Secrets dienen zur eindeutigen Identifizierung der jeweiligen Tokens und ßießen auch in die Signierung der Nachrichten mit ein. OAuth-Anfragen können zwei Identifier haben: *oauth_consumer_key* und *oauth_token*. Damit der Server die Authentizität der Anfragen überprüfen kann, muss der Client beweisen Eigentümer des Identifiers zu sein. Dies wird dadurch erreicht, dass jeder Identifier ein korrespondierendes Passwort bzw. Shared Secret besitzt.

Im Wesentlichen besteht OAuth aus drei Hauptschritten: 1. Der Consumer fragt ein unbestätigtes Anfrage-Token und Anfrage-Token-Geheimnis beim Provider an. 2. Der Benutzer bestätigt das Anfrage-Token durch eine Authentifizierung und eine einfache Zustimmung. 3. Der Consumer fragt mit dem bestätigten Anfrage-Token ein Access-Token und Access-Token-Geheimnis beim Provider an. Diese drei Hauptschritte (siehe Abbildung 2) werden nachfolgend genauer erläutert. Zur Kommunikation stehen drei Varianten zur Auswahl: Der Client kann den HTTP-Header *Authorization* zur Kommunikation nutzen und fügt in diesen alle notwendigen OAuth-Parameter durch Kommas separiert hinzu (die nachfolgenden Beispiele nutzen diese Methode). Das Auth-Schema wird hierbei auf "OAuth" gesetzt. Die Parameter können aber auch per POST im HTTP-Anfragebody in "application/x-www-form-urlencoded"-Form übertragen werden, sowie als Querystring in einer Get-Anfrage.

*1) Abrufen eines Anfrage-Tokens:* Für die nachfolgenden Erklärungen wird das vorherige Beispiel[3] mit Jane, Müller und flickr wiederaufgenommen. flickr bietet einige Endpoit URLs an (z.B. "photos.flickr.com/initiate"), über die der Client per HTTP-Post oder HTTP-GET kommunizieren kann. Weitere Endpoit URLs, die in diesem Beispiel verwendet werden, sind "photos.flickr.com/authorize" und "photos.flickr.com/token".

Jane besucht die Webseite von Müller.de und wird aufgefordert Zugriff auf ihre bei flickr hochgeladenen Bilder zu gewähren. Zuerst muss ein temporärer Identifier erzeugt werden, der die anstehende Delegation eindeutig identifiziert. Dazu wird in Schritt A (siehe Abbildung 2) zuerst ein Anfrage-Token von flickr geholt. Hierfür muss Müller.de folgende Anfrage an flickr schicken:

```
POST /initiate HTTP/1.1
Host: photos.flickr.com
Authorization: OAuth
 realm="Photos",
 oauth_consumer_key="dpf43f3p2l4k3l03",
 oauth_signature_method="HMAC-SHA1",
 oauth_timestamp="137131200",
 oauth_nonce="wIjqoS",
 oauth_callback="http%3A%2F%2Fdrucker.
 mueller.de%2Fready",
 oauth_signature="74KNZJeDHnMBp0EMJ9ZHt
```

[3]Das Beispiel ist aus dem OAuth RFC 5849 [13] vom April 2010 entnommen.

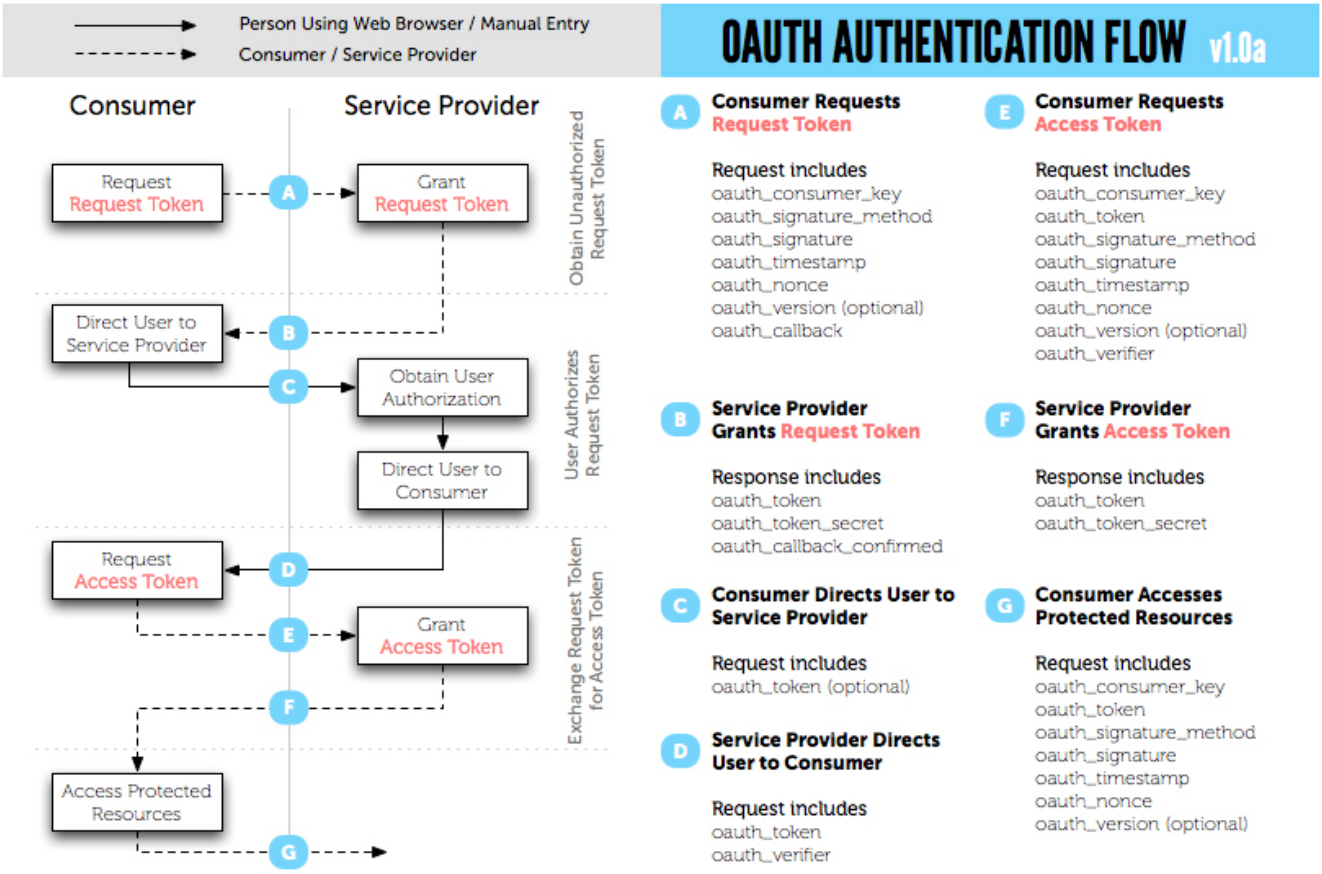


Figure 2. Protokollablauf einer Autorisierung mittels OAuth. Der Autorisierungsprozess besteht aus den Schritten A-G. Die rechte Hälfte der Grafik listet die Parameter der Get-Anfragen bzw. die Paramter in den Antwortnachrichten auf. Die gestrichelten Linien trennen den Prozess in die drei Hauptbestandteile auf. Quelle: [12]

```
%2FXKycU%3D"
```

Das OAuth-Protokoll stellt drei Signatur-Methoden zur Verfügung: “HMAC-SHA1”, “RSA-SHA1” und “PLAINTEXT”. Für die Erstellung der Signatur werden bestimmte [13] HTTP-Header bzw. OAuth-Parameter zu einem Basisstring konkateniert, der wiederum zusammen mit einem Key in die Eingabe für die Signatur-Algorithmen fließt. Der Key ist spezifiziert als “Consumer Secret” + “&” + “Token Secret”. Weil die Token Secrets erst in späteren Schritten zur Verfügung stehen, wird dieses im ersten Schritt durch einen Leer-String substituiert (Key = “Consumer Secret” + “&”). Das Ergebnis der Signatur wird im base64-enkodierten Feld *oauth_signature* übermittelt. Das Feld *oauth_callback* enthält eine URL, an die der Benutzer nach Zugriffseinwilligung weitergeleitet wird. Ist die Anfrage valide, so antwortet flickr mit einem unbestätigten Anfrage-Token:

```
HTTP/1.1 200 OK
Content-Type: application/x-www-form-urlencoded
oauth_token=hh5s93j4hdidpola&
oauth_token_secret=hdhd0244k9j7ao03&
oauth_expires_in=3600&
oauth_callback_confirmed=true
```

In der Antwort sind die so genannten temporären Credentials enthalten, die die anstehende Autorisation einem Benutzer zuordenen. Das Feld *oauth_token_secret* stellt das bereits erwähnte Anfrage-Geheimnis dar und ist ein temporäres Geheimnis mit dem der temporäre Identifier *oauth_token* geprüft werden kann. Da diese Antwort bereits sensible Informationen enthält, muss der Server eine sichere Verbindung mit TLS oder SSL erzwingen.

*2) Weiterleitung des Benutzers zum Provider und dortige Zugriffseinwilligung:* Nach Erhalt des unbestätigten Anfrage-Tokens, muss dieses von Jane bestätigt werden. Dazu wird Jane mit dem unbestätigten Token auf die Seite “https://photos.flickr.com/authorize?oauth_token=hh5s93j4hdidpola” weitergeleitet (siehe Schritt B in Abbildung 2). flickr fordert Jane auf sich mit Benutzername und Passwort anzumelden und anschließend eine Zustimmung zu erteilen, dass “drucker.mueller.de” auf ihre Fotos zugreifen darf. Jane bestätigt die Aufforderung und wird dann an die Adresse, die in *oauth_callback* angegeben wurde, weitergeleitet (siehe Schritt D). Die Weiterleitungsadresse sieht dann wie folgt aus:

```
http://drucker.mueller.de/ready?
oauth_token=hh5s93j4hdidpola&
oauth_verifier=hfdp7dh39dks9884
```

Der Schlüssel *oauth_verifier* repräsentiert hierbei das bestätigte Token.

*3) Anfrage nach einem Zugriffs-Token:* Durch die Rückleitung an "drucker.mueller.de/ready..." weiß Müller.de, dass Jane eingewilligt hat und kann mit dem Protokoll fortfahren. Als nächstes fordert Müller.de mit dem bestätigten Request-Token über eine sichere Https-Verbindung ein Access-Token an. Die Anfrage nach einem Zugriffs-Token wird verschickt:

```
POST /token HTTP/1.1
Host: photos.example.net
Authorization: OAuth
 realm="Photos",
 oauth_consumer_key="dpf43f3p2l4k3l03",
 oauth_token="hh5s93j4hdidpola",
 oauth_signature_method="HMAC-SHA1",
 oauth_timestamp="137131201",
 oauth_nonce="walatlh",
 oauth_verifier="hfdp7dh39dks9884",
 oauth_signature="gKgrFCywp7rO0OXSjdot%2FI
 HF7IU%3D"
```

flickr.com validiert die Anfrage und antwortet mit einem Access-Token:

```
HTTP/1.1 200 OK
Content-Type:application/x-www-form-urlencoded
oauth_token=nnch734_normallymuchlonger&
oauth_token_secret=pfkkdhi9sl3r4s00&
oauth_expires_in=3600
```

*4) Access protected Resources:* Mit dem Access-Token kann Müller.de nun auf geschützte Ressourcen von Jane zugreifen:

```
GET /photos?file=vacation.jpg&
size=original HTTP/1.1
Host: photos.example.net
Authorization: OAuth
 realm="Photos",
 oauth_consumer_key="dpf43f3p2l4k3l03",
 oauth_token="nnch734_normallymuchlonger",
 oauth_signature_method="HMAC-SHA1",
 oauth_timestamp="137131202",
 oauth_nonce="chapoH",
 oauth_signature="MdpQcU8iPSUjWoN%2FUDMsK2
 sui9I%3D"
```

Der Client kann nun solange Fotos anfordern, bis das Zugriffs-Token abläuft oder solange Jane bei flickr.com angemeldet ist. Beim Yahoo OAuth wird bei obiger Antwortnachricht noch das Feld *oauth_session_handle* zurückgeliefert. Mit diesem Wert lässt sich ein dauerhaftes Access-Token bei Yahoo anfordern. Generell ist es notwendig sich in die Dokumentationen des jeweiligen Providers einzulesen, da sich die Endpoint URLs und deren jeweilige Aufrufe unterscheiden können.

### C. Sicherheitstechnische Gesichtspunkte

Die dreibeinige OAuth Core 1.0 Spezifikation [12] hat eine gravierende Sicherheitslücke namens *Session Fixation Attack* [14] und ist deshalb seit dem 24. Juni 2009 veraltet. Die OAuth Core 1.0 Revision A Spezifikation [15] gleicht diese Schwachstelle wieder aus. Diese ist allerdings auch veraltet und Entwickler sollten sich an RFC 5849 [13] halten. Die Durchführung des Angriffs ist relativ simpel. Der Angreifer meldet sich bei einem legitimen Consumer *X* an und startet die Autorisierung. Der normale Protokollablauf wird nach Schritt B "Grant Access Token" (siehe Abbildung 2) unterbrochen, indem der Angreifer nicht der Weiterleitung auf "https:// photos.flickr.com/-authorize?oauth_token=hh5s93j4hdidpola" zur Bestätigung des Anfrage-Tokens folgt. Der Angreifer speichert sich diese URI und bringt ein Opfer zum Beispiel durch Phishing, Cross-Site-Scripting (XSS) etc. dazu diesen Link zu klicken. Wenn das Opfer auf den Link klickt, wird er aufgefordert den vom Angreifer gestarteten Autorisierungsprozess zu vervollständigen. Man beachte, dass das Opfer auf die originale Bestätigungsseite des Service Providers geleitet wird, um einen legitimen Consumer zu bestätigen. Das Opfer hat keine Möglichkeit den Angriff zu erkennen, wenn es beabsichtigt den Consumer *X* zu autorisieren. Nachdem das Opfer den Zugriff gewährt hat, kann der Angreifer das bestätigte Request-Token benutzen, um den Protokollablauf zu vervollständigen und ein Zugriffstoken anfordern. Wenn der Angreifer einen Account beim Consumer *X* hat, könnte die Autorisation über längere Zeit bestehen bleiben. Dieser Angriff funktioniert, weil das Anfrage-Token eine Assoziation zwischen Benutzer und Autorisationsanfrage herstellt – in diesem Szenario aber eine Assoziation mit einem böswilligen Benutzer. Der Consumer wird das erhaltene Autorisationstoken folglich dem böswilligen Benutzer zusprechen.

## IV. Zusammenfassung

OpenID und OAuth sind zwei Protokolle, die man als Informatiker bereits mindestens einmal gesehen oder gehört hat. Jedoch wissen nur Wenige genau, für was diese Protokolle eingesetzt werden oder wie sie funktionieren. Facebook hat seit Mitte 2009 seine Tore für Third-Party Identifier geöffent [16]. Das heißt es ist nun möglich sich bei Google Mail einzuloggen und sich beim ersten Aufruf der Facebook-Seite automatisch anmelden zu lassen. Dies setzt natürlich voraus, dass man seinen Google-Account mit dem Facebook-Account verknüpft hat. Für große Unternehmen wie Yahoo oder Microsoft hat es einen entscheiden Vorteil selbst als OpenID Provider zu agieren. Sie können dadurch das Surfverhalten ihrer Nutzer besser tracken, da bei jeder Anmeldung Kontakt mit dem Provider aufgenommen werden muss. So verwundert es nicht, dass man sich mittlerweile sogar mit seiner Windows Live ID bei Relying Parties anmelden kann. OpenID hat durch seine weite Verbreitung das Potential zur ersten Single-Sign-On Lösung für das gesamte Web zu werden. In Zukunft könnte man sich also nur noch fragen, "bin ich im Netz eingeloggt oder bin ich es nicht". Ein weiterer Vorteil ist ein einheitlicher Anmeldevorgang auf allen unterstützen Webseiten. Man muss sich natürlich bewusst machen, dass der OpenID-Provider der Single-Point-Of-Failure ist.

Das Web 2.0 bietet eine Vielfalt an Inhalten, die von Benutzern erzeugt werden. Möchte man diese geschützten Inhalte in anderen Webservices wiederverwenden, bedarf es einer Authorisierungsfunktion. Das Authorisierungsprotokoll OAuth wurde in der Arbeit klar von OpenID differenziert und dessen Funktionsweise geschildert. Der Authentifizierungspart im OAuth-Protokoll kann auch sehr gut mit OpenID realisiert werden. Um auf die Daten und Funktionen von Google,

Twitter und Yahoo zugreifen zu können, kommt man unlängst nicht mehr an OAuth vorbei.

## REFERENCES


[1] "OpenID Foundation website." [Online:accessed 18-December-2011]. [Online]. Available: http://openid.net/

[2] OAuth, "OAuth Community Site." [Online:accessed 18-December-2011]. [Online]. Available: http://oauth.net/

[3] "Meistbesuchte Websites in Deutschland - Reichweite im September 2011 - Statistik." [Online:accessed 18-December-2011]. [Online].Available: http://de.statista.com/statistik/daten/studie/161592/umfrage/meistbesuchte-websites-in-deutschland-nach-reichweite/

[4] "Windows Live ID." [Online:accessed 18-December-2011]. [Online]. Available: https://accountservices.passport.net/ppnetworkhome.srf?vv=1200&mkt=DE-DE&lc=1031

[5] Jayant Gandhi, "OpenID, A Spam Heaven?" [Online:accessed 18-December-2011] [Online].Available: http://www.jkg.in/103-openid-a-spam-heaven/

[6] heise Developer, "Identity Management: Authentifizierungsdienste mit OpenID." [Online:accessed 18-December-2011] [Online]. Available: http://www.heise.de/developer/artikel/ Identity-Management-Authentifizierungsdienste-mit-OpenID-227202.html

[7] "URI Generic Syntax.", RFC 3986 [Online:accessed 18-December-2011] [Online]. Available: http://www.ietf.org/rfc/rfc3986.txt

[8] "Extensible Resource Identifier (XRI) Resolution Version 2.0." [Online:accessed 18-December-2011] [Online]. Available: http://docs.oasis-open.org/xri/2.0/specs/cd02/xri-resolution-V2.0-cd-02.pdf

[9] "Yadis Specification Version 1.0." [Online:accessed 18-December-2011] [Online]. Available: http://yadis.org/papers/yadis-v1.0.pdf

[10] "Diffie-Hellman Key Agreement Method.", RFC 2631 [Online:accessed 18-December-2011] [Online]. Available: http://www.ietf.org/rfc/rfc2631.txt

[11] "XML-Signature Syntax and Processing.", RFC 3275 [Online:accessed 18-December-2011] [Online]. Available: http://www.ietf.org/rfc/rfc3275.txt

[12] "OAuth Core 1.0." [Online:accessed 18-December-2011] [Online]. Available: http://oauth.net/core/1.0/#authstep1

[13] Eran Hammer-Lahav, "The OAuth 1.0 Protocol.", RFC 5849 [Online:accessed 18-December-2011] [Online].Available: http://tools.ietf.org/html/rfc5849

[14] OAuth, "OAuth Security Advisory: 2009.1." [Online:accessed 18-December-2011] [Online]. Available: http://oauth.net/advisories/2009-1/

[15] "OAuth Core 1.0a." [Online:accessed 18-December-2011] [Online]. Available: http://oauth.net/core/1.0a/

[16] "Facebook's transparent use of OpenID." [Online:accessed 18-December-2011] [Online]. Available: http://joss.blogs.lincoln.ac.uk/2009/06/16/facebooks-transparent-use-of-openid/

# GSM/3G/4G/DECT Security

Dominik Deuter
Ulm University
Ulm, Germany
Email: dominik.deuter@uni-ulm.de

***Abstract*—This paper deals with the issue of security in mobile networks. Initially a generation-overview of mobile communication standards is given. It describes the GSM and DECT technology with the help of technical basics as well as security features discussed and suggestions given for security flaws. Also security aspects of 3rd and 4th generation are listed. In addition a security comparison between generations is realized.**

## I. Introduction

In recent years the number of users of mobile networks has developed exponentially. Shown in Figure 1. According to a study of the United Nations, mathematically every second cosmopolitan uses a cell phone. The most important growth regions for the mobile providers are the developing countries. In the industrial countries, however the market penetration in mobile network is over 100 percent because many people own more than one cell phone. The origin idea of mobile phoning has rapidly developed. New functions were implemented. SMS services and other data transmissions for example video broadcasting are no longer imagine without.

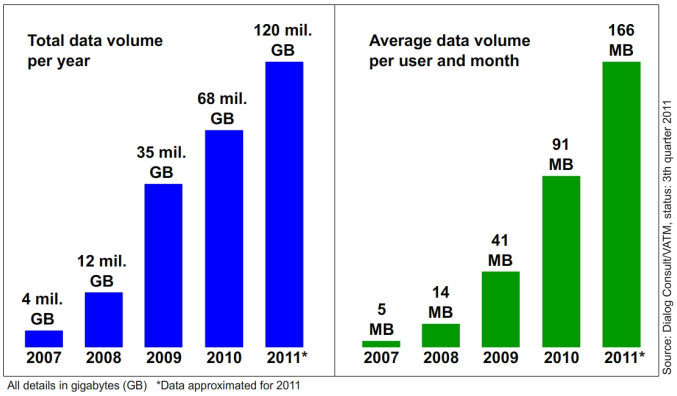


Figure 1. Data volume in mobile network [1]

In research, new mobile communication technologies are constantly developed, which are becoming more complex. Thus the aspect of security in such systems should be considered especially. This paper is organized as follows: In section II the historical development of mobile networks will be elucidated. At this a classification of communication technologies in different generations will be done. Section III deals with GSM system. Introductory started with technical basics. Following security aspects will be discussed. It will be shown security flaws as well as counteractive measures against unauthorized access in GSM systems. In section IV security aspects of 3rd generation followed by comparison of GSM, which is associated with 2nd generation will be considered. Section V points an overview of 4th generation. At this architecture and security aspects of LTE are shown. In section VI DECT system will be analyzed. At this technical basics and security aspects will be discussed. Section VII gives the conclusion.

## II. Historical development of mobile networks

### A. 1G - first generation

The 1st generation (1G) of mobile networks emerged in the 1980s. The most successful standards of this generation were *Nordic Mobile Telephone* (NMT), *Total Access Communications System* (TACS) and *Advanced Mobile Phone Service* (AMPS). For the first time, NMT enabled to perform international roaming. This allowed that even outside its home network via other providers (communications service provider) to make calls and data exchange. Even though we are in 3rd generation of mobile networks nowadays, part of 1st generation is still in use. Sooner or later 1st generation is shelved to make way for new generations in frequency range. In germany it is complete disabled.

### B. 2G - second generation

The 2nd generation (2G) came during 1990s. In contrast to 1G data transfer was converted from analog to digital. Due a synchronous division of the frequency spectrum by a temporal or code identified division more subscribers can use the same frequency, resulting in a considerable efficiency. Four standards of second generation have prevailed: the *Digital Advanced Mobile Phone System* (D-AMPS); *Code-Division Multiple Access IS-95* (CDMA); *Personal Digital Cellular* (PDC) plus the most successful and widespread system of 2G the *Global System for Mobile Communications* (GSM). Originally developed as a European standard it reached quickly worldwide use except of North America. In North America *Personal Communication System-1900* (PCS-1900) prevailed. It is mentionable in context of second generation systems the digital cordless system *Digital Enhanced Cordless Telecommunications* (in short DECT). This system consists of a base station and a group of subscribers. Because DECT has only a limited range it is mainly capable for local use. Additional mobile technologies of that generation are HSCSD, GPRS and EDGE.

### C. 3G - third generation

In the same year as GSM was launched commercially, the *European Telecommunications Standards Institute* (ETSI) started the standardization of 3rd generation (3G) by mobile telecommunications networks. The *Universal Mobile Telecommunications System* shortly UMTS. By the way ETSI is a member of the global *3rd Generation Partnership Project* (3GPP) which cares for standardization in mobile radio. For data transmission in mobile telecommunications networks of 3G a so-called *Wideband Code Division Multiple Access* multiplexing (WCDMA) is applied. In this connection a frequency bandwidth of 5 MHz is used. This allows a data rate of up to 384 kbit/s (compared to GSM 9.6 kbit/s). Thereby it made possible to realize video communication. Data rates of different techniques, see Figure 2. [2]

### D. 4G - fourth generation

In 2005 due the project *Next Generation Mobile Networks* (NGMN) a standardization of 4th generation (4G) began which inserted commercially since 2010. A technology of this generation is the so-called *Long Term Evolution* (LTE) which is the first uniform mobile communications standard at all. Due LTE bandwidths reach up to 1 gbit/s. This means that in near future the classical broadband connection could be replaced by it. [3]

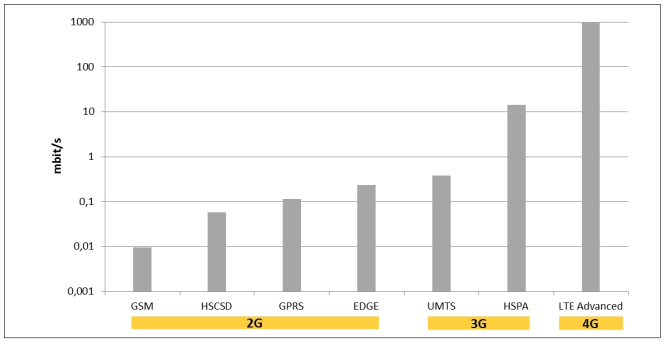


Figure 2. Data rates [4]

## III. GSM-System

In the 1980s, in Europe there were too many national and incompatible radio networks. For this reason, the idea of a standardized pan-European cellular radio network was born. For the development a working group was founded. The so-called *Group Speciale Mobile* (GSM). Today, the acronym GSM stands for *Global System for Mobile Communications* to underline the claim of a worldwide standard. Goals of GSM are:

- A wide range of voice and data service offerings,
- Compatibility with wired networks (ISDN, telephone network, data network) by means of standardized interfaces,
- International subscriber-access to system,
- Automatic pan-European roaming and handover,
- High efficiency in utilization of the frequency spectrum,
- Support for various types of mobile devices (e.g. vehicle, portable and handheld mobile phones),
- Digital transmission,
- Independence of manufacturers,
- Low costs for infrastructure and devices.

### A. The architecture of GSM

The GSM system is divided into three subsystems: (see Figure 3)

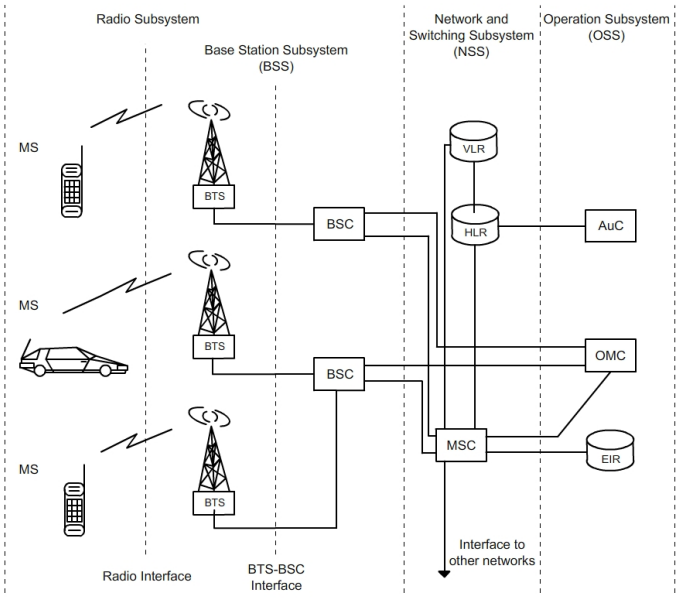


Figure 3. GSM architecture [5]

*1) Radio Subsystem (RSS):* RSS consists of *mobile stations* (MS) and *base station subsystems* (BSS). The MS contains two parts. The radio interface for hardware and software components and the *Subscriber Identity Module* (SIM), which contains all subscriber-specific information. Via SIM, all subscribers are uniquely identified in a mobile network. Furthermore a MS inside GSM network assigned following numbers and identities:

- International Mobile Station Identity (IMSI)
- Temporary Mobile Station Identity (TMSI)
- Mobile Station International ISDN Number (MSISDN)
- Mobile Station Roaming Number (MSRN)

BSS comprises whole radio-related part of GSM network. These include the component *Base Transceiver Station* (BTS), which consists of transmission and reception systems and also signal processing. Depending on type of antenna, one or more cells can be supplied. As well as the *Base Station Controller* (BSC). BSC is responsible for managing the radio interface. It handles reservation and disengaging of radio channels as well as handover management. A handover ensures a smooth transition of a connection between the changes of two cells. In general a BSC manages several BTS and is connected via MSC with NSS.

*2) Network and Switching Subsystem (NSS):* NSS forms a transition-network between radio network and public networks such as telephone and data network. The *Network and Switching Subsystem* makes functions available whose suitable realization and implementation is task of the manufacturer and network operator.
The *Mobile Services Switching Center* (MSC) is a digital switching center that performs switching functions and manages the network. Furthermore, the MSC is the connector between mobile and wired networks. Further tasks include all signaling processes that are required for establishing, terminating and managing connections. Including services like call forwarding, call barring, teleconference and charge for the called subscriber.
The *Home Location Register* (HLR) is the database where all information of a mobile phone subscriber is stored. There are both permanent data e.g. the phone number as well as temporary data like current location stored. Each mobile phone subscriber is registered exactly in a home register where also charge data are recorded.
The *Visitor Location Register* (VLR) provides as well as HLR a database that is assigned with an MSC. It manages the subscribers who are currently located within the responsibility of this MSC. VLR contains a copy of necessary information by HLR. HLR is determined by the call number of the mobile station, in which master data of subscriber are located. The *Home Location Register* returns the information in which VLR the subscriber is located and allows thereby a connection with MSC. The VLR thus controls the allocation of roaming numbers.

*3) Operation Subsystem (OSS):* OSS is responsible for subscriber's administration. Subscriber data stored in HLR enable network operators to offer mobile rates and bill the subscriber. In addition, the OSS includes functions of network operation and maintenance. So it is possible to provide the operator monitoring, control and intervention options.
The *Operation and Maintenance Center* (OMC) controls and monitors as a central core other network elements. OMC also calculate statistical data about status and utilization of network elements.
The *Authentication Center* (AuC) contains all information which is served to protect subscribers identity. AuC provides a mechanism which will secure mobile communication against eavesdroping. The use of the radio interface is only possible if the subscriber is authorized. Because the radio interface is generally vulnerable against access, special arrangements (e.g. assignment of an authentication key for each subscriber and encryption of transmitted information) were taken to prevent abuse. The authentication algorithm and encryption key are stored in AuC (explanation of security aspects in the next section). The *Equipment Identity Register* (EIR) is a central database that stores the *International Mobile Equipment Identity* (IMEI). It contains a white, black and gray list. The white list contains the IMEIs of valid mobile stations, the black list contains all IMEIs of stolen or blocked mobile stations and the gray list contains the IMEIs of devices with a dysfunction wherefore no services are provided for them. [5]

On the successor systems HSCSD, GPRS and EDGE which allow higher data rates for example (see Figure 2) are not discussed in this paper. For further information to these techniques see [6].

## B. Security

By the architecture of GSM, it is possible to implement security services such as anonymity, authentification and confidentiality with user data. Starting point is the SIM card. It contains all necessary information for a network access. The SIM card itself is secured from unauthorized access by a self-selected *Personal Identification Number* (in short PIN). On it, the secret, randomly generated 128-bit *individual subscriber authentication key* (Ki) and the 15-digit IMSI are stored.

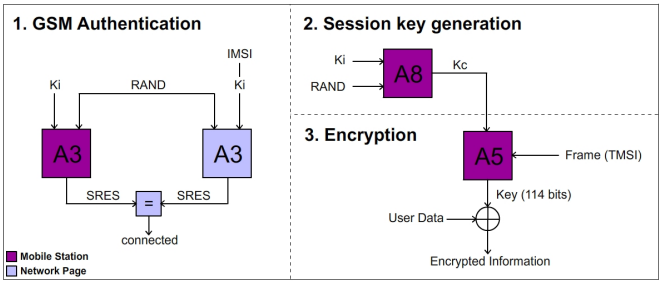


Figure 4. GSM authentification, session key generation and encryption [7]

During the authentication the subscriber is assigned to a random number RAND which is generated in the AuC. From RAND calculates the mobile station via Ki and A3 the authenticator *SRES (signed response) = Ki(RAND)*. The result is sent to VLR and compared with the value which is generated there. If these values match, the subscriber is authenticated. All transmitted messages of the subscriber are being protected. For this purpose a random number RAND, the authentication key (Ki) and the A8 algorithm generates a transmission key (Kc).
Due to these and the application of A5 algorithm a 114-bit long security key is generated. To ensure the confidentiality of transmitted information of the mobile subscriber, the *Temporary Mobile Station Identity* (TMSI) is changed periodically by using the TDMA method. This will prevent that a cellular connection can be allocated by eavesdropping a particular mobile subscriber. The TMSI is always assigned temporarily by the current VLR and encrypted transmitted at the mobile station. After that the subscriber's data XOR with the generated key (the receiver can decrypt the text with help of the key) and creates the encrypted information. [5][7]

*1) Security flaws in GSM:* Since GSM system was designed with an over-the-air encryption it is very vulnerable to any third party. The following security flaws includes the GSM system:

- Unilateral authentication
  In GSM there is only one-sided authentication. The system checks each subscriber to his identity but not every subscriber the system. Thus it is possible to

attackers to use a false BTS on the same network access code. Thus, the attacker can authenticate and perform a man-in-the-middle attack. In this case, exchanging data can be manipulated or created. This is the basis for the use of IMSI-catchers. These are devices that simulate a base station. This allows to read the IMSI and isolate the location of the mobile station.
In development of GSM such a scenario was considered for not possible because equipment would have been too expensive. Nowadays just a fraction of cost is required for this. Therefore it is used by security authorities to detect criminal activity and by the criminals.

- Implementation of A3/A8 algorithm
  The GSM architecture provides manufacturers the opportunity to use any algorithm for the A3/A8 implementation. But most of them took advantage of COMP128 which was developed by the GSM Association. In COMP128 several security flaws were uncovered.
  *Example: Decryption of subscribers authentication key (Ki).* Of the transmission key Kc are set 10 bits to zero from right and the encryption algorithm is 1024-times weaker and more susceptible due the reduced key space.

- Cloning the SIM card
  The *Smartcard Developer Association* and the ISAAC research group found a vulnerability in COMP128 that allows to disclose the Ki on SIM. Thus the attacker can intercept phone calls and text messages in the name of the subscriber. However, the attacker can not circumvent the implementation of GSM network. The system detects if idential SIM are connected at the same time. Thereby the system locks the account.

- Protection of short distance
  Only the distance between MS and BTS is encrypted. The other parts of the network are unprotected and the information is transmitted unencrypted. This is one of the main flaws of GSM especially if the communication between BTS and BSC is wireless.

- Visibility of the encryption
  The encryption is controlled by the BTS. The user is not informed if the encryption mode is disabled in BTS. Thus the subscriber data will be transmitted unencrypted to another mobile station. [7]

*2) Elimination of the security flaws:* In 2002, many efforts have been made to develop new encryption algorithms for GSM. These were implemented in A3/A8 to prevent cloning the SIM card. This had the disadvantage that new SIM card and the software of HLR had to be revised. These algorithms are further developments of COMP128-1 called COMP128-2 and COMP128-3. This now prevents cloning the SIM card as well as modify the Ki. This should then be used in both the BTS and the mobile station. Only changing the encryption algorithms is not enough to ensure the security. It is still possible that the attacker pretends to be a trusted network and induce mobile stations to disable the encryption mode. Thus it is also necessary to change the authentication protocols.
If an operator wishes to use other security algorithms it is just possible in consultation the involved software and hardware manufacturers.
A further improvment is the encryption of the backbone in GSM (see *protection of short distance*). So it is more difficult to intercept transmitted data. [7]

## IV. 3G ASPECTS

After describing the basics and security aspects of GSM this section gives an overview of the *Universal Mobile Telecommunications System* (in short UMTS) which corresponds the 3G of mobile networks.
UMTS is backward compatible to GSM extension systems. This has the advantage that mobile operators do not have too much overhead in upgrading. The main change from UMTS to GSM is the new radio transmission technology which needs a new base station system. Furthermore, the frequency band of a UMTS system with 5 MHz is much more expanded than in GSM system (200 kHz). This enables a higher user capacity. The switching subsystem is up to software updates for the MSCs largely preserved. The UMTS system includes an asymmetrical data transmission (data transmission on multiple channels). This enables video telephony because picture and sound to be transmitted simultaneously. [8]

### A. Security

The security mechanisms of UMTS are similar to GSM. However, they were strong revised and improved. Mobile stations are now using USIM cards (*Universal Subscriber Identity Module*) which represent an extended feature of the SIM card.
To define a system of 3G certain security properties are preconditioned. These are divided into five overarching category groups:

1) Network access security
2) Security within the network area
3) Security in the user area
4) Security within the scope
5) Visibility and configurability of security

*1) Network access security:* This area deals with secure access to the services of 3G.

- Confidentiality of subscriber identity
  The *International Mobile Subscriber Identity* (IMSI) of the mobile station must be secure against eavesdropping during the radio transmission. Otherwise, inferences to location or the subscribers personal data can be drawn.

- Bilateral authentication
  It is necessary that the mobile station authenticates to the dialing network and the network authenticates to the

mobile station.

- Confidentiality of data
  It must be ensured that the data generated by subscriber are not intercepted due to encryption algorithms.

- Data integrity
  The mobile station and the network have to be able to verify received data to the consigner and origin.

*2) Security within the network area:* The components within the network area need to authenticate each other to ensure the integrity and confidentiality of subscriber's data.

*3) Security in the user area:* This section is responsible for security in the mobile station. To be able to access the USIM, the user has to authenticate himself. This is done by using a secret (e.g. PIN). Similarly, the USIM must authenticate to the mobile station which in turn is done by knowledge of a secret.

*4) Security within the scope:* There must be a secure exchange of data between application of the mobile station and the provider's area. For this purpose an encryption of data in the radio connection and an encryption in the whole network area is necessary, therefore, a complete point-to-point encryption.

*5) Visibility and configurability of security:* The subscriber needs to have control about activated and disabled services. The following security features are defined:

- Visibility
  The subscriber must be informed if a service uses an encrypted or unencrypted connection. Likewise if a subscriber enters a cell with other safety requirements. For example, the change from a UMTS to a GMS cell.

- Configurability
  The subscriber can decide whether to authenticate the UMTS card is to be carried out or whether he wants do admit or accept unencrypted calls. [9]

### B. Safety comparison between 3G systems and GSM

Many security aspects and security flaws that were not reflected or fixed in GSM has been improved by third generation. In this section, the UMTS system using the GSM system is tested for security aspects. These are listed below:

- It is necessary a bilateral authentication between the mobile station and network now. In GSM this was unilateral.

- The key lengths are now consistently set to 128 bits. In GSM the keys had different lengths. The transfer key Kc in GSM had only 64 bits which could be decrypted relatively easily.

- The data are consistently encrypted point-to-point. In contrast to GSM where only the radio interface between mobile station and BTS is encrypted.

- The subscriber is always be informed about the used encryption. [10]

## V. 4G Overview

The following 4th generation of mobile networks intended to support multimedia applications in its entirety. This requires a very high data rate. There are currently two technologies that are considered for a 4G mobile network. The *Long Time Evolution* (LTE) and the *Worldwide Interoperability for Microwave Access* (WiMAX). 4G is based in contrast to previous generations on a TCP/IP architecture.
In this paper the 4G aspects are considered largely on the LTE system. The first section describes the architecture of LTE. Important components that constitute the LTE system are listed. Following the security aspects are considered.

### A. LTE architecture

The basic scheme of the UMTS architecture is also maintained in the LTE architecture (see Figure 5). This enables a cost-effective retrofit with the providers. Furthermore, significantly fewer components in the architecture involved than this was the case in older mobile networks. This allows a much lower transmission time.

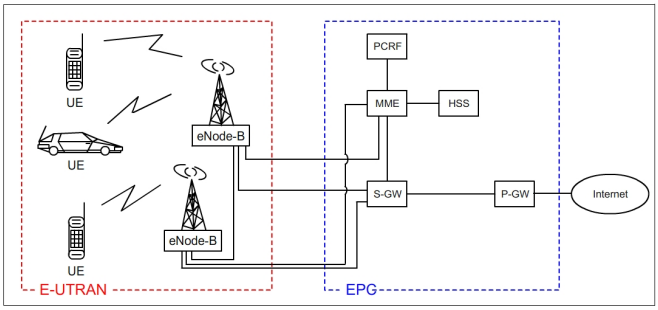


Figure 5. LTE architecture [11] [12]

The mobile stations are referred to as *User Equipment* UE. The base stations get the designation *eNode-B* and are connected to perform a fast handover. These components form the radio network *Envolved UMTS Terrestrial Radio Access Network* (E-UTRAN). The UEs connect (mostly over fiber optic lines) over the eNodes with the so-called *Envolved Packed Core* (EPC) which represents the core network. This is completely IP based. The *Management Mobility Entity* (MME) is responsible for authentication and location of the UEs. The MME access to the *Home Subscriber Service* (HSS) which is the database of the system. In HSS all information needed for authentication is stored. Thereafter the MME allocate a *Serving-Gateway* (S-GW) to the subscriber. The S-GW is now responsible for routing and transmitting the packets.

Furthermore, the core network includes a so-called *Policy and Charging Rules Function* (PCRF) which determines the services of a specific tariff.
About the *Packet Data Network Gateway* (P-GW) finaly, an Internet connection is established and an IP address is assigned to the UE. [11] [12]

### B. Security

In 4G LTE was further security aspects considered than was the case in older mobile networks. For example different levels of abstraction for the *identification number* (ID) were equipped. In 2G systems static IDs were used on the SIM card. In 4G these are dynamic now. Another aspect is the increased security in the transmission between UE and MME. This result in the requirements which would be useful for a security system of 4G: increased robustness compared to 3G, confidential handling of personal subscriber data, a solid inspection at the processes of subscriber authentication and network as well as maintaining of security on transition to other radio networks.
Compared to 3G security in LTE has an advanced authentification and key provisions, a more complex key hierarchy and more complex security between network components.
In relation to the key hierarchy, LTE provides five different keys. These relate to various uses and their validity. All keys derived from the so-called *K Key*. This key is permanently stored on the USIM of the UE. The five different keys are described more detailed in [10]. The 4th generation of mobile networks is still underutilized in commercial use (2010/2011). Thus many security flaws are still largely unknown. [11]

## VI. DECT SYSTEM

The *Digital Enhanced Cordless Telecommunications* (DECT) was adopted as a standard by the *European Telecommunications Standards Institute* (ETSI) in 1992. In Nort America this technology will be deployed in a different frequency spectrum.
The DECT standard is a digital mobile network which is especially characterized by its high quality of voice and data transmission. Since 2006 there is an extension of the DECT standard, the *CAT-iq* which is not discussed in this paper. For more information about *CAT-iq* see [13]. The DECT technology is used primarily in office buildings and home areas. The range, in comparison to other mobile network technologies is relatively low. It has outdoor a maximum of 300 meters (328 yards). DECT is mainly used for telephone communications but finds for example use in radio transmission of alarm systems.

### A. The architecture of DECT

With DECT complete cordless private subarea systems (e.g. the communication of departments in firms) can be constructed. Beside the connection to a public telephone network it is possible to build up an internal communication link between multiple mobile devices free of charge via a DECT base station. For the following explanation of architecture see Figure 6.

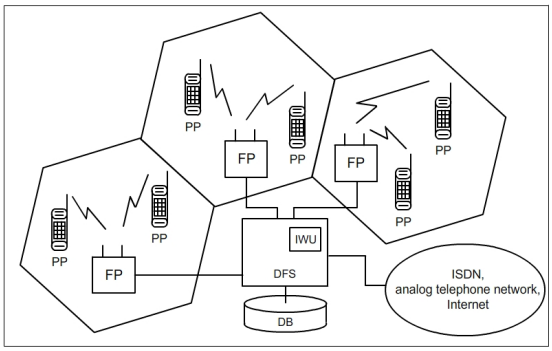


Figure 6. DECT architecture [13]

Thus, the DECT system consists of a fixed station, the *Fixed Part* (FP) and multiple mobile stations, the *Portable Parts* (PP). Furthermore, there is a direct mode that allows two PPs can directly communicate with each other.
DECT is capable of multi-cells. This means that a DECT network consists of several FPs and is connected to a central switching component *DECT Fixed System* (DFS). The covered area can be provided coverage by overlapping radio cells. The database of DFS includes the subscriber and terminal device management. For connection to an external telephone network there is in DFS a *Interworking Unit* which ensures for the adjustment of DECT-specific protocols to the protocols of the analog telephone network or ISDN. For smaller DECT-system structures for example in the home area the functions of DFS are implemented in the FP.

When setting up a connection every FP sends a signal in the form of 40 bits long *Radio Fixed Part Identity* (RFPI) which is a globally unique number. Thereby the PPs it's possible to decode RFPI and to identify FPs located in the area. At PPs is information about the respective access rights, so-called PARKs (*Portable Access Rights Keys*) is stored. The access rights are agreed during a *subscription*, that is upon registration of a PP with an FP. As a result, the PP has now the services of the FP (and the associated network).A PP always choose the FP for which exist the access rights and the strongest signal.
During inscription of a PP, the PP receives from FP a unique temporary number. The 20 bit *Temporary User Identity* (TPUI). With this the PP is proclaimed during an incoming call.

### B. Security

Because DECT is a radio-based method, must be attached a lot of importance to security. Otherwise, it is possible to eavesdrop or interfere the communication. In this section specific security mechanisms, identified potential vulnerabilities and possible countermeasures are discussed.

*1) Security measures:*

- Authentication
  Basically, an authentication should take place between all components of a system. Thus, each PP must authenticate before connecting. This is based on a so-called *challenge-response method*.
  The FP is sent to the PP a 64 bit random number RAND and RS. The PP is now calculated using a long-time secret UAK (*User Authentication Key)* which must be stored in both devices and RS by using algorithm A11 a key intermediate KS. KS generates with the random number RAND by using algorithm A12 the response *RES1* (32 bit). The PP will now transfer RES1 to FP. Now FP compares RES1 with the previously generated value XRES1 in FP. If these values match, PP is authenticated.
  The authentication of the FPs compared with the PP is optional. The procedure is similar to the authentication of the PP. However, the tasks are reversed.
  First PP generates a 64 bit random number RAND and sends it to the FP. The FP generates the value RS (64 bit), thus using the UAK and algorithm A21 and generate a 128 bit intermediate key KS. From that is then generated with the random number RAND and the algorithm A22 the response RES2 (32 bit). These values are sent to the PP which performs the same calculation. If these two calculations agree, the FP is authenticated.

- Encryption
  In DECT it is optionally possible to encrypt the subscriber data via the radio interface. The encryption is based on a stream cipher. The *key stream* is generated by a key stream generator and a 64 bit long encryption key CK.
  The key stream is then combined XOR with the subscriber data. For decryption is now on the receiving end the key stream thus contains the subscriber data with the help of CK and XOR converted into plain text. The key stream generator is the so-called *DECT Standard Cipher* (DSC). This algorithm is not published, however partially reconstructed. Furthermore, to criticize in DECT that encryption is optional.

- Authentication of subscriber
  For changes of important settings there is a 4 to 8 digit secret number called PIN which to be entered by the subscriber.

*2) Security flaws:* Many security flaws result from weaknesses in the protocols. Following are some security flaws in DECT are discused.

- Insecure default settings
  In many cases default settings are configured by the manufacturer to use weak. Often the delivery status of the device for example, the PIN value is *0000*. This allows hackers easily to get access to the device.

- Interference of the DECT infrastructure
  By using of jammers it is possible to restrict the availability of the DECT infrastructure.

- Eavesdropping
  With DECT devices it is possible to eavesdrop on room conversations. This is particularly possible with devices that support a so-called *baby monitor function*.

*3) Protection measures:* In this section, protection measures are presented which should make the DECT system more secure.
It is useful to check before using the default settings of the device. It is appropriate not to use trivial PINs and if the option to *enforce an encryption* is possible then it should be enabled.
It should be avoided the use of *DECT repeaters* which amplify the signal if these require a deactivation of encryption. Instead, additional FPs should be used.

In summary, DECT offers a higher security standard than other cordless standards which is sufficient to use a radio scanner to eavesdrop a call. However many devices on the market where encryption is not implemented. This can be achieved by a special equipment or special software to eavesdrop communications over DECT relatively easy. [13]

## VII. Conclusion

In this paper an overview of systems of different generations was given. Mainly the architecture and the security aspects were picked out. Potential weaknesses or protective measures were identified. Currently we are in terms of time at the transition between 3rd and 4th generation. It will be interesting what the future will bring in development of 4G techniques. By the global networking of these systems it will be increasingly important to take account to appropriate security measures in the future.

## References


[1] "Datenaufkommen," Computerbild 24/11, p. 8, 2011.
[2] Juha Korhonen, *Introduction to 3G Mobile Communications*, 2. ed. Artech House, Boston 2003, pp. 1 – 12.
[3] "Overview 4th generation," Accessed: Nov 20th, 2011. [Online]. Available: http://www.itwissen.info/definition/lexikon/fourth-generation-4G-4-Generation.html
[4] "Data rates," Accessed: Nov 21th, 2011. [Online]. Available: http://www.elektronik-kompendium.de/sites/kom/0406221.htm
[5] Bernhard Walke, *Mobilfunknetze und ihre Protokolle 1*, 3. ed. Teubner, Stuttgart/Leipzig/Wiesbaden 2001, pp. 135 – 151; 325 – 327.
[6] Neeli R. Prasad, "GSM Evolution towards Third Generation UMTS/IMT2000." Lucent Technologies Zadelstede 1-10, Nieuwegein, The Netherlands.
[7] Ali Asghar Beheshti Shirazi, Mohsen Toorani, "Solutions to the GSM Security Weaknesses." Iran University of Science and Technology; Department of Electrical Engineering, Teheran 2008.
[8] Erik Pischel, "Architektur des Universal Mobile Telecommunications Systems." Seminar paper; Humbold-Universität zu Berlin; Institut für Informatik, Berlin 2002, p. 15.
[9] Stefan Kralicek, "UMTS Security." Seminar paper; Ruhr-Universität Bochum; Lehrstuhl Kommunikationssicherheit, Bochum 2002, pp. 14 – 16.

[10] Arne Wolf Kösling, "Sicherheitsanalyse in drahtlosen Netzen." Seminar paper; Carl-von-Ossietzky-Universität Oldenburg; Fachbereich Informatik; Abt. Rechnernetze und Telekommunikation, Oldenburg 2002, pp. 68 – 72.

[11] N. Seddigh, B. Nandy et al., "Security Advances and Challenges in 4G Wireless Networks." Eighth Annual International Conference on Privacy, Security and Trust, Ottawa 2010.

[12] "LTE architecture," Accessed: Dec 8th, 2011. [Online]. Available: http://www.elektronik-kompendium.de/sites/kom/1608191.htm

[13] Heinz Gerwing, Jörg Mattke et al., "Drahtlose Kommunikationssysteme und ihre Sicherheitsaspekte." Bundesamt für Sicherheit in der Informationstechnik, Bonn 2009.

# RFID Security and Privacy

Philipp Hock
Institut für Medieninformatik
Universität Ulm
Ulm, Deutschland
Email: philipp.hock@uni-ulm.de

***Abstract*—RFID-Tags sind mittlerweile in vielen Alltagsgegenständen integriert. Die Sicherheits- und Privatsphärerisiken werden dabei oftmals vernachlässigt. In diesem Artikel wird zunächst ein Überblick über allgemeine Probleme zum Thema Sicherheit und Privatsphäre in RFID Systemen geschaffen. Im weiteren Verlauf wird auf mögliche und bereits existierende Mechanismen eingegangen, die Sicherheit und Privatsphäre verbessern können und wo deren Schwachstellen sind. Außerdem werden Authentisierungsmechanismen für verschiedene RFID-Tagklassen genauer betrachtet.**

## I. Einleitung

RFID steht für *radio-frequency identification* und bezeichnet eine Technik zur drahtlosen und automatischen Identifizierung von Gegenständen. Die Interaktion findet in der Regel zwischen einem sogenannten *RFID-Tag* und einem Lesegerät statt, wobei das Lesegerät auch als Sender dient, welches Daten auf das Tag schreiben kann. Die vom Lesegerät ausgesendete elektromagnetische Strahlung versorgt in den meisten Fällen den Tag mit Strom.

Die simpelste Form von RFID-Tags sendet lediglich einen Identifikationsstring an ein interagierendes Lesegerät, welcher eine Id und weitere Metainformationen enthält. Es findet also keine wechselseitige Kommunikation, sondern ein unilateraler Informationsaustausch statt, wie man es beispielsweise von Barcodes kennt. Jedoch enthalten Barcodes gewöhnlich nur Informationen, wie: "Das ist ein Produkt $Y$ der Marke $X$", während RFID-Tags Informationen zur eindeutigen Bestimmung des Tags senden: "Das ist ein Produkt $Y$ der Marke $X$ mit der Id $Z$".

RFID kann aber viel mehr, als man es bisher von Barcodes kennt. Anwendungsszenarien reichen vom Türöffner bis zum bargeldlosen Zahlungsmittel. Waschmaschinen, die erkennen welchen Waschgang die enthaltene Wäsche braucht oder Mikrowellen, die von alleine die richtige Garzeit festlegen, sind durchaus realistische Szenarien. Komplexe Protokolle und Berechnungen sind ebenfalls mit RFID-Tags möglich, wenn auch nicht alle Tags die selben technischen Möglichkeiten haben. Sind diese Informationen für jedes Lesegerät auslesund änderbar, entstehen Probleme im Bereich Sicherheit und Privatsphäre. Besonders die drahtlose und automatisierte Interaktion ist ein Merkmal von RFID. Werden beispielsweise Codes für einen Türöffner unverschlüsselt übertragen, könnten diese von Lesegeräten in unmittelbarer Umgebung unbemerkt ausgelesen werden. Geben die Tags zusätzlich Produktinformationen preis, können mitgeführte Gegenstände von fremden Lesegeräten ausgelesen werden.

Dieser Artikel befasst sich mit den Problemen bei RFID-Systemen und welche Mechanismen es ermöglichen, die Sicherheit und Privatsphäre zu schützen. In Kapitel II wird ein Überblick über die Technik und Klassifizierungen verschiedener RFID-Tags gegeben. Kapitel III erläutert verschiedene Aspekte der Sicherheit und Privatsphäre. In Kapitel IV werden verschiedene Authentisierungsmechanismen genauer erläutert.

## II. Technik und Klassifizierungen

RFID-Tags bestehen in der Regel aus einem Chip und einer Antenne. Der Chip hat einen analogen Sendeschaltkreis und einen digitalen Informations- und Verarbeitsunsschaltkreis. Bei der Kommunikation werden Hochfrequente Radiowellen übertragen, die das Tag über seine Antenne aufnimmt. Dabei ist die Antenne einerseits die Stromversorgung, indem das Strahlungsfeld des Lesegeräts die Energie bereitstellt, andererseits werden über die Antenne auch Nutzdaten übertragen. RFID-Tag bezeichnet also die Kombination aus Chip und Antenne, währen bei RFID-Chip der eigentliche Chip gemeint ist. Im Folgenden sind die Unterschiede beider Begriffe jedoch belanglos und beziehen sich auf das Selbe.

Nach Juels [1] lassen sich RFID-Tags bezüglich ihrer Stromversorgung kategorisieren:

- Passive Tags werden allein durch das elektromagnetische Wechselfeld des Senders mit Strom versorgt.
- Semi-Passive Tags werden durch das Feld aktiviert, haben dann aber eine eigene Stromversorgung.
- Aktive Tags dagegen besitzen eine statische Stromversorgung und können auch Initiator einer Kommunikation sein.

Die Stromversorgung spielt eine große Rolle bei der Funktionalität der Chips. Ein aktiver Chip beispielsweise kann leistungsfähigere Hardware besitzen, um viele kryptographische Operationen durchzuführen, die ein passiver Chip nicht durchführen kann. Sie werden zum Beispiel bei der Logistik in Frachtschiffkontainern eingesetzt. Passive Tags sind jedoch billiger und meist kleiner. Sie finden immer mehr Verwendung als Ersatz von Barcodes, sind aber auch in Kreditkarten oder Ausweisen enthalten. [1]

Manche Chips werden sogar unter die Haut implantiert. Anfangs wurden nur Haustiere mit solchen Chips ausgestattet. In Diskotheken beispielsweise werden mittlerweile aber auch beim Menschen kleine Chips als Bargeldersatz unter die Haupt implantiert.

Auch die Einteilung in die kryptographische Mächtigkeit ist eine Möglichkeit der Klassifizierung, allerdings ist der Übergang hier eher fließend. Grob können hier die Tags in funktionale Merkmale eingeteilt werden:

- Tags ohne kryptographische Funktionen.
- Tags mit schwachen kryptographischen Operationen, wie Pseudoszufallszahlengenerator oder Hashing.
- Tags, die symmetrische Verschlüsselungsverfahren implementieren (wie AES).
- Tags, die asymmetrische Verschlüsselungsverfahren implementieren (wie RSA).

Werden die Nachrichten nicht vom Tag generiert, sondern nur ausgegeben, können die Daten bereits signiert oder verschlüsselt im Tag abgelegt sein. Dann muss das Tag keine besonderen kryptografischen Funktionen besitzen. Verschlüsselung und Signierung setzen also nicht zwangsläufig eine stärkere kryptografische Funktionalität des Chips voraus, da diese in manchen Fällen auch vom Leser übernommen werden können.

Besonders relevant sind die sogenannten *low-cost Tags*. Das sind passive Tags, die bereits in vielen Gegenständen enthalten sind, wie zum Beispiel auf Produkten im Supermarkt, in Kreditkarten, Personalausweisen oder Schlüsseln. Es ist abzusehen, dass es in Zukunft immer mehr Produkte geben wird, die einen passiven RFID-Chip enthalten.

Der *EPC Class-1 Gen-2* Chip [2] ist ein weit verbreiteter Chip, der als Produktcode von Waren verwendet wird. Ari Juels [1] zeigt auf, dass der EPC-Chip weder ausreichend Schutz der Privatsphäre bietet, noch über Sicherheitsmechanismen gegen das unerlaubte Klonen des Tags verfügt.

Ebenfalls weit verbreitet ist der *Mifare Classic* [3]. Er ist in vielen Studentenausweisen integriert und dient der Identifikation, als Zugangskontrolle und als Zahlungsmittel in vielen Universitäten. Garcia et al. [4] zeigen, dass dessen *Crypto-1* Verschlüsselung leicht geknackt werden kann. Dieser bietet ebenfalls keinen Schutz der Privatsphäre oder Mechanismen gegen das Manipulieren oder Klonen der Daten auf dem Chip.

## III. Sicherheit und Privatsphäre

Zunächst muss geklärt werden, was man im Zusammenhang mit RFID unter Privatsphäre und Sicherheit versteht und welche grundsätzlichen Szenarien denkbar sind.

### A. Sicherheit

Die Sicherheit von RFID-Systemen lässt sich grob in drei Kategorien unterordnen: Authentisierung, Manipulation und Übertragung. Bei der Authentisierung geht es darum, sicher zu stellen, dass ein Gerät berechtigt ist, Informationen vom Kommunikationspartner zu erhalten oder zu senden. Hierbei kann man weiter unterscheiden:
Bei der *Leser-zu-Tag-Authentisierung* wird sichergestellt, dass sich ein Lesegerät beim Tag authentifiziert.
Die *Tag-zu-Leser-Authentisierung* garantiert, dass ein Lesegerät ein bestimmtes Tag oder eine Gruppe von Tags erkennt. Kombiniert ergibt sich die *Gegenseitige-Authentisierung*, bei der sichergestellt wird, dass Tag und Lesegerät berechtigt sind, miteinander zu kommunizieren.

Schutz vor Manipulation der Nutzdaten ist beispielsweise bei Reise- und Personalausweisen notwendig. Tags lassen sich auf verschiedene Arten vor der Manipulation enthaltener Daten schützen. Eine relativ sichere Methode ist das Signieren der Daten. Dabei wird zusätzlich ein Verifikationsstring $v$ im Tag gespeichert, wobei $v$ der verschlüsselte Hashwert der Daten ist. Damit kann über die Signatur geprüft werden, ob die Daten des Tags gültig sind. Nur wer den Signaturschlüssel kennt, kann die Daten ändern, ohne sie ungültig zu machen. Alternativ könnten die Nutzdaten auch komplett verschlüsselt auf dem Chip gespeichert werden. In der Kryptologie ist Verschlüsselung allerdings kein Verfahren, dass Authentizität gewährleistet. Ein Angreifer könnte beispielsweise willkürlich Datenfragmente des Chips ändern, in der Hoffnung wieder einen gültigen, aber veränderten Datensatz zu erhalten.

Der Schlüssel birgt ein Sicherheitsrisiko, denn ist der geheime Schlüssel erst einmal bekannt, ist das komplette System nicht mehr sicher. Bei *low-cost Tags* werden sensible Daten, wie zum Beispiel Kontostände, häufig in externen Datenbanken gehalten. Hierbei kann der Chip eine eindeutige Id auf einen Datenbankeintrag besitzen. Es muss aber auch hier sichergestellt werden, dass die Id des Chips nicht einfach kopiert werden kann. Das Signieren der Id bietet zusätzlichen Schutz vor Manipulation. Um allerdings Schutz vor dem Klonen des Tags zu bieten, müssen komplexere Funktionen, wie Authentisierung, im Chip implementiert werden.

Sind zwei Kommunikationspartner gegenseitig authentisiert, besteht das Problem, dass der darauf folgende Datenaustausch weiterhin abgehört oder sogar kompromittiert werden kann. Das Abhören von Nachrichten kann nur durch Verschlüsselung verhindert werden. Um Nachrichten lediglich vor Fälschung zu sichern, können sie mittels Signierung oder Verschlüsselung geschützt werden. Ob eine verschlüsselte Übertragung sinnvoller ist, als eine signierte, hängt vom Anwendungsfall ab. Digitale Signaturen werden allerdings mittels asymmetrischen Verschlüsselungsalgorithmen erzeugt, daher ist die Anforderungen an die Hardware in beiden Fällen sehr hoch. Liegt der Rechenaufwand beim Lesegerät, spielt die hohe Hardwareanforderung in der Regel keine Rolle.

Sogenannte *covert channels* sind ein weiteres Problem. Ist die Kommunikation zwischen Leser und Tag verschlüsselt, hat auch der Besitzer des Tags keine Kontrolle über die übermittelten Daten. Ari Juels beschreibt dies treffend [5]:

> But if you have a wireless identification device implanted in your body, and it's emitting cryptographically protected values, how do you know that all it's doing is identifying you? How do you know it's not acting as a sensor and secretly reporting medical information to your employer [...]

Die Lösung dieses Problems erweist sich als nicht trivial und wird von Baily et al. 2007 [6] genauer erläutern.

### B. Privatsphäre

Um die Privatsphäre des Besitzers zu schützen, muss sichergestellt werden, dass ein Tag nur gewisse Informationen an Lesegeräte sendet, die dafür berechtigt sind. Dabei unterscheidet man einerseits zwischen Informationen, die das

Identifizieren von Personen ermöglichen und andererseits das Auflisten von Gegenständen, die eine Person bei sich hat. Man spricht hier von Lokalisierung und Inventarisierung von Personen. Der *EPC Class-1 Gen-2 Chip* enthält sowohl eine eindeutige Seriennummer, als auch Produktinformationen, die von jedem Lesegerät ausgelesen werden können. Selbst wenn man mit Seriennummern alleine keine Rückschlüsse auf eine bestimmte Person ziehen kann, können Identifikatoren von Tags mit Namen kombiniert werden, beispielsweise über Kreditkarteninformationen [5].

Die Lesereichweite spielt eine große Rolle, in Bezug auf Privatsphäre. Nach Ari Juels [1] gibt es nicht nur eine Lesereichweite für Tags, sondern vier:
DIE NOMINELLE REICHWEITE ist die Entfernung, die ein Lesegerät mit normaler Antenne haben muss, um das Tag mit Strom zu versorgen und auslesen zu können.
DIE ERWEITERTE REICHWEITE erhöht die nominelle Reichweite mittels stärkeren Antennen oder Antennenarrays.
DIE TAG-ZU-LESER-ABHÖRREICHWEITE ist die Entfernung, die ein Lesegerät zum Auslesen eines Tags haben muss, wenn dieses bereits von einem anderen Lesegerät mit Strom versorgt wird.
DIE LESER-ZU-TAG-ABHÖRREICHWEITE ist die Entfernung, die ein Lesegerät haben muss, um Daten zu einem RFID-Tag senden zu können. Dies können theoretisch mehrere Kilometer sein.

Der *EPC Class-1 Gen-2* kann die Leser-zu-Tag-Abhörreichweite durch einen einfachen Mechanismus verringern: Dabei sendet das Tag zuerst einen zufälligen Bitstring $B$. Die Nutzdaten $N$ werden vom Leser XOR verschlüsselt zurückgesendet: $N = B \oplus N$. Dadurch können Lesegeräte nur Daten senden, die sich mindestens in *Tag-zu-Leser-Abhörreichweite* befinden. Der einfachste Schutz vor unberechtigtem Lesezugriff, ist eine physische Abschirmung, beispielsweise durch Metallfolien.

Die Entfernung zwischen Tag und Leser kann aber auch als Maß für die Vertraulichkeit des Lesers sein. Ari Juels [1] zeigt, dass einfache *signal-to-noise-ratio* Messungen mit wenig Aufwand im Chip approximativ die Entfernung bestimmen können. Anhand dessen kann eine Policystrategie anwenden werden. Beispielsweise geben Tags bei weit entfernten Anfragen nur wenig oder keine Informationen preis, bei geringer Entfernung jedoch sehr viel. Das Problem der Tag-zu-Leser-Abhörreichweite besteht jedoch auch hier, sodass ein entfernter Leser die Kommunikation von zwei naheliegenden Geräten mithören kann.

Ein weiterer Mechanismus, um Privatsphäre sicherzustellen, ist im *EPC Class-1 Gen-2* eingebaut. Dieser besitzt einen Selbstzerstörungsmechanismus, den man mit einem 32 Bit langen Codewort auslösen kann. Bei einem Warenkorb im Supermarkt z.B. könnte man diesen Auslösen, sobald die Ware bezahlt worden ist. Das Verfahren macht den Chip allerdings auch für praktische Anwendungen nach dem Kauf unbrauchbar.

Juels [1] schlägt daher eine Erweiterung vor, bei dem sich das Tag nur temporär deaktivieren lässt und durch bestimmte Berührung oder das Senden eines Aktivierungsschlüssels wieder aktiviert wird. Nachteile sind in beiden Fällen schnell ersichtlich: Bei einem Aktivierungspasswort liegt die Passwortverwaltung beim Menschen, der sich ohnehin schon schwer tut, seine Passwörter zu verwalten. Physischer Kontakt dagegen macht sämtliche Vorteile von automatischer, drahtloser Übertragung bei RFID zunichte.

Ein anderer Kontrollmechanismus wäre ein von Rieback et al. [7] vorgeschlagener RFID-Guardian. Dieser übernimmt die Aufgaben, ähnlich einer Firewall, um Lesegeräten den Zugriff zu gewähren oder zu verweigern. Dieser könnte den Zugang zum Tag aktiv blockieren oder bei Bedarf das Tag aktivieren, beziehungsweise deaktivieren. Integriert in ein NFC[1] fähiges Smartphone mit GPS und Internetzugang, wäre es ein mächtiges Werkzeug. Allerdings müsste das Smartphone dann ständig für Überwachungsaufgaben bereit sein. Besonders problematisch ist die Zugangskontrolle. Das Gerät müsste den Zugang zum Tag ständig überwachen und freigeben, wenn ein Leser dazu berechtigt ist.

Sogenannte *Blocker Tags*, ohne die aufwendige Guardianfunktionen, würden möglicherweise bereits ausreichen, um ein zufriedenstellendes Maß an Privatsphäre zu gewährleisten. Juels et al. [8] schlagen eine passive Lösung in Form eines Tags vor, die das Antikollisionsprotokoll (Singulation) ausnutzt, um solch einen Blockiermechanismus zu ermöglichen. Das Protokoll ist in Abbildung 1 skizziert. Kurz gesagt kann so ein Tag Kollisionsnachrichten verschicken, die ein Lesegerät daran hindert, Tags auszulesen, die sich in der Nähe des Blockertags befinden. Das Blocken würde beispielsweise bei Tags geschehen, die ein bestimmtes, sogenanntes "Privacy Bit" gesetzt haben. Der Mechanismus basiert darauf, dass ein Lesegerät mit mehreren Tags gleichzeitig interagieren kann. Beim Singulation Protokoll fragt der Leser sequenziell jede Stelle des Identifikators aller Tags ab. Diese Antworten also immer mit Null oder Eins. Die Identifikatoren spannen so einen Binärbaum auf. Dadurch können die Identifikatoren aller Tags ermittelt werden. Haben zwei Tags unterschiedliche Bits an einer Stelle, kollidieren die Signale. Sendet ein Tag immer eine kollidierende Nachricht bestehend aus Null und Eins, können die Identifikatoren der Tags nicht mehr bestimmt werden. Das Privacy Bit ist in dem Fall die erste Stelle des Identifikators.

Ein denkbares Szenario wäre beim Einkaufen im Supermarkt. Einkaufstüten oder ähnliches könnten mit einem Blocker-Tag ausgestattet sein. Das Privacy Bit würde beim Bezahlen bei den Tags der Waren gesetzt werden. So lassen sich zwar keine Informationen beim Transport auslesen, könnten aber zu Hause in Kombination mit einem intelligenten Kühlschrank nützliche Informationen austauschen. Der Nachteil ist, dass der Ansatz nur für ein bestimmtes Protokoll und für spezielle Tags (Privacy Bit) funktioniert.

Problematisch ist hier, dass ein Leser alle möglichen Kombinationen von Identifikatoren durchprobiert, wenn ein Blockertag mit ihm interagiert. Dies sind unter Umständen viele Millionen Möglichkeiten.

[1] Near Field Communication ist ein Übertragungsstandard, der die Kommunikation mit RFID-Tags ermöglicht.

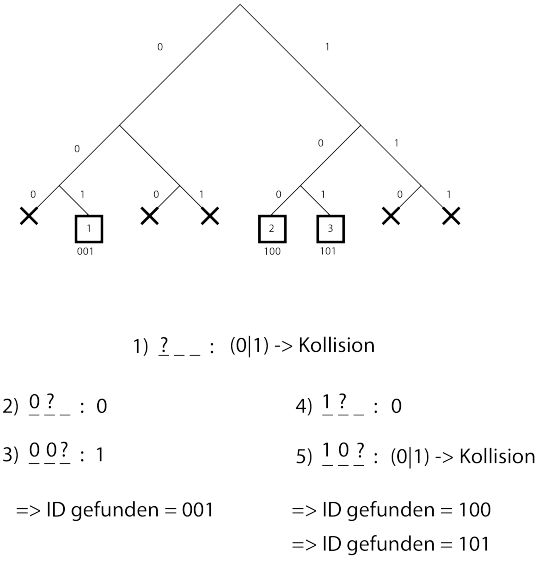


Abbildung 1. Das Singulation Protokoll: Ein Leser fragt sequentiell alle Stellen der Ids von RFID-Tags ab. Die Tags antworten mit der jeweiligen Stelle. Übermittelt ein Tag eine Null, ein anderer eine Eins, tritt eine Kollision auf. Der Leser weiß, es gibt mehrere Tags mit unterschiedlichen IDs und fährt rekursiv mit der Suche fort.

## IV. Authentisierung im Detail

Die meisten Probleme im Zusammenhang mit Privatsphäre und Sicherheit können durch entsprechende Authentisierungsmechanismen gelöst werden. Zugang zu einem Tag oder Lesegerät wird nur dann gewährt, wenn dieser autorisiert ist. Gemeinsamkeit fast aller Authentisierungsmechanismen ist das Überprüfen eines sogenannten *shared secrets*, also ein Geheimnis oder Passwort, das von beiden Authentisierungspartnern verifiziert werden kann. Ausnahme ist hier nur das asymmetrische Verfahren, welches mit öffentlichen und privaten Schlüsselpaaren arbeitet.

RFID-Tags sind durch ihre unterschiedliche Rechenleistung nicht mit einheitlichen Protokollen ausstattbar. Für unterschiedliche Szenarien gibt es unterschiedliche Authentisierungsprotokolle.

Der naive Ansatz zur Authentisierung ist das einfache Übertragen des Geheimnisses. Dies kann aber von fremden Lesegeräten abgehört werden und muss zusätzlich verschlüsselt oder anderweitig verschleiert werden. Dazu kann das Passwort gehashed versendet werden. Der Empfänger gleicht den empfangenen Hashwert mit den Werten in einer Datenbank ab, stimmen sie überein, ist die Authentisierung erfolgreich. Um zu verhindern, dass ein Angreifer einfach diesen Hashwert ausliest und sendet, kann das Protokoll zu einem Challenge-Response Protokoll erweitert werden, ähnlich der *Digest Access Authentication* [2] im HTTP/1.1 Protokoll. Der zu authentisierende muss eine zufällige *Nonce*[3] mit einhashen, die er vom Authentisierungspartner bekommt, was dazu führt, dass das Passwort immer einen anderen Hashwert hat. Diese Methode bildet jedoch ein Risiko der Privatsphäre des Authentisierenden, falls ein aktiver Leser die selbe Nonce versendet, erhält er auch immer den gleichen Hashwert und kann damit eventuell Rückschlüsse auf die Person ziehen. Dies könnte man verhindern, indem man an das Passwort noch einen zufälligen Bitstring (Salt) vor dem Hashen anhängt, dann den Hash und zusätzlich noch den Salt versendet.

Ein Challenge-Response Verfahren verhindert zum Einen das Klonen des Tags, zum Anderen bietet es ausreichend Schutz der Privatsphäre. Lediglich durch physischen Zugang zum Tag ist es möglich, ein Tag zu Klonen (Side-Channel Attacke[4]). Das Verfahren wäre in einem Tag einfach zu implementieren, lediglich ein Pseudozufallszahlengenerator und der Hashalgorithmus sind nötig.

Hat jedes Tag seinen eigenen Schlüssel (shared secret), müssen alle gültigen Schlüssel in einer Datenbank im System hinterlegt werden. Bei mehreren Datenbankeinträgen muss bei der Authentisierung auch jeder Eintrag verglichen werden, was das Authentisieren bei großen Datenbanken verlangsamt. Die Überprüfung geschieht folgendermaßen: Das System sendet eine Nonce an den Tag, der einen Bitstring $hk = salt|hash(key|nonce|salt)$ zurücksendet. Um zu Überprüfen, ob der Schlüssel korrekt ist, muss das System $hk$ mit allen möglichen gültigen Werten vergleichen:

```
def is_validKey(key, nonce, salt):
    for valid_key in DataBase:
        valid_bits =
              hash(valid_key+nonce+salt)

        if valid_bits == valid_key:
            return True;

    return False;
```

Dies benötigt eine Laufzeit von $\mathcal{O}(n)$, wobei $n$ die Anzahl der möglichen Authentisierungsschlüssel ist. Gängige Mechanismen bilden die Id auf ein Passwort ab, um die Suche zu beschleunigen. Jedoch wäre die Id wieder ein eindeutiger Identifikator. Mit ihm ließe sich das Lokalisieren von Personen ermöglicht.

Ein einfacher Ansatz, die Id nicht im Klartext zu übertragen, wäre eine verschlüsselte Authentisierung. Dies könnte realisiert werden, indem beide Authentisierungspartner einen gemeinsamen Schlüssel $k$ besitzen. Die Kommunikation bei der Authentisierung wird mit $k$ verschlüsselt. Das oben genannte Challenge-Response Verfahren ist auch hier die Grundlage der Authentisierung, allerdings kann die Id mit übertragen werden, ohne die Privatsphäre zu gefährden, denn die Kommunikation erfolgt verschlüsselt. Das System entschlüsselt die Nachricht und kann dann über die Id schnell prüfen, ob der übertragene Hashswert korrekt ist. Eine sichere Kommunikation kann ebenfalls gewährleisten werden, indem der gesamte Datenverkehr verschlüsselt erfolgt. Dies benötigt allerdings leistungsfähigere Hardware. Angriffspunkt ist hier $k$. Ist $k$ bekannt, kann die Übertragung abgehört werden.

[2] http://www.ietf.org/rfc/rfc2617.txt

[3] Eine Nonce ist ein zufälliger Bitstring, der einmalig verwendet wird, um zum Beispiel Replay-Attacken zu verhindern.

[4] Kryptoanalytische Methode, bei der die physische Implementierung des Systems angegriffen wird [9].

Asymmetrische Verfahren funktionieren nach einem ähnlichen Prinzip. Lediglich sind zum Ver- und Entschlüsseln unterschiedliche Schlüssel notwendig. Gegenseitige kryptografische Authentisierung bzw. Leser-zu-Tag-Authentisierung erfordert im Tag allerdings ein hohes Maß an Rechenleistung, die bisher nur teure, oft aktive, Chips erbringen können.

In den Abschnitten IV-A und IV-B werden zwei Ansätze vorgestellt um die Suche nach dem Schlüssel zu beschleunigen, während die Privatsphäre geschützt bleibt. In Abschnitt IV-C und IV-D wird auf zwei alternative, minimalistische Verfahren eingegangen. Abschnitt IV-E und IV-F gehen genauer auf Verfahren ein, die stärkere kryptografische Operationen nutzen, um Authentisierungen zu ermöglichen.

### A. Authentisierung durch Baumstrukturen

Molnar und Wagner [10] beschreiben ein Verfahren, das die Suche nach dem richtigen Authentisierungsschlüssel beschleunigt. Hierbei hat jedes Tag mehrere Schlüssel, die in einer hierarchischen Baumstruktur angeordnet sind. In Abbildung 2 hat Tag 1 beispielsweise die Schlüssel [5,2,1]. In reellen Szenarien sind die Schlüssel nicht nur ein oder zwei Stellen lang, sondern mehrere hundert Bits. Der Leser authentisiert zunächst den Wurzelknoten, eliminiert dann in jedem Schritt (im Fall eines Binärbaums) einen Zweig im Baum. Der vollständige Schlüssel eines Tags entspricht dann einem Pfad des Baumes von der Wurzel bis zum Blatt. Der Vorgang ist vereinfacht in Abbildung 2 veranschaulicht. Durch die Tiefensuche der Schlüssel ergibt sich eine Laufzeit von $\mathcal{O}(\log n)$. Der Nachteil ist, ändern sich Schlüssel im System, muss der gesamte Baum geupdated werden.

### B. Counter

Juels [1] schlägt eine weitere Methode vor, um die Laufzeit der Suche zu reduzieren. Grundlage ist ein Zähler im Tag, der bei jeder erfolgreichen Authentisierung inkrementiert wird. Das System kennt für jeden Tag den ungefähren Wert des Zählers. (Im Folgenden soll zur vereinfachten Darstellung angenommen werden, dass das System den genauen Wert kennt. Die Suche wird dabei nur um einen Faktor verändert, der in der asymptotische Laufzeit keine Rolle spielt.)

Das System hat eine Tabelle mit Zählerwerten $C$ als Schlüssel und den dazugehörigen Passwörtern. Zum Beispiel: [4711] = [$key_0$,$key_1$,..,$key_n$]. Die Suche wird beschleunigt, indem das Tag bei der ersten Abfrage des Lesers den Wert des Zählers übermittelt. Der Leser muss dann nur noch für die Passwörter des Schlüssels $C$ den Hashwert berechnen. Einzige Einschränkung ist hier, dass alle Lesegeräte im System vernetzt sein müssen. Durch ungeschickte Verteilung können die Zählerwerte allerdings auch eine Art Id sein, was Aufschlüsse über die Identität des Tags geben könnte. Sind alle Schlüssel des Systems mit ähnlichen Zählerwerten belegt, bringt das Verfahren keine Geschwindigkeitsvorteile.

### C. pseudonym throttling

Ein Ansatz von Ari Juels von 2005 [11] ist ein Verfahren mit minimalistischen kryptografischen Berechnungen. Es eignet

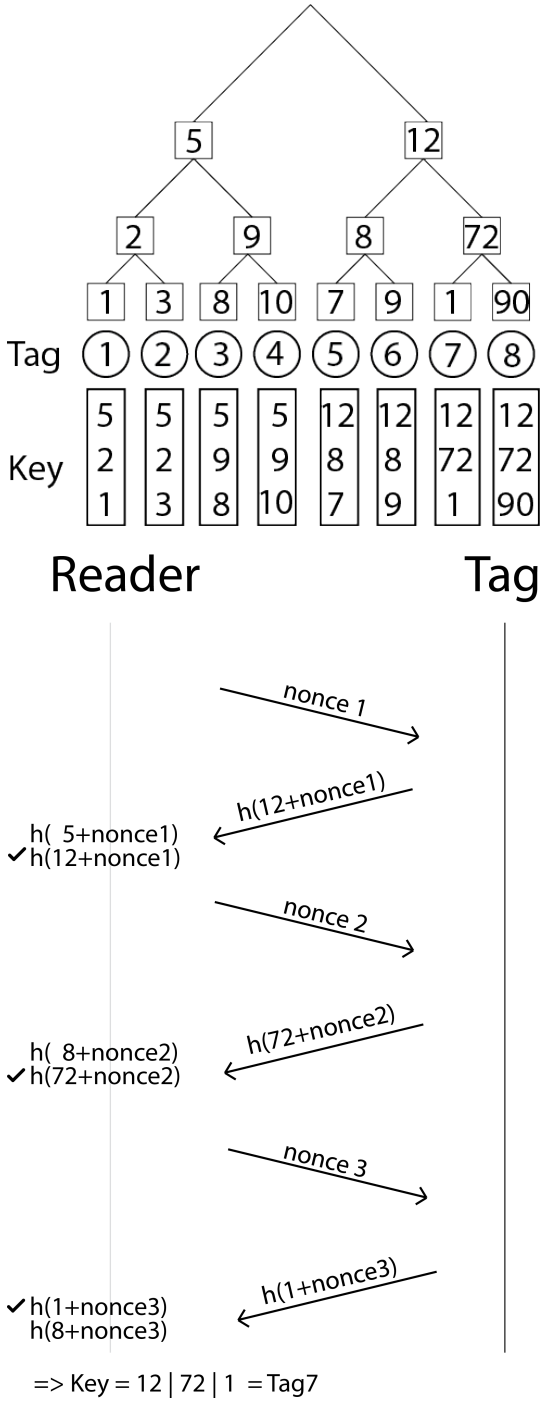


Abbildung 2. Authentisierung mittels Baumstruktur.

sich ebenfalls für *low-cost Tags*. Das Verfahren berücksichtigt, dass Angriffsszenarien bei RFID-Tags nur unter verschiedenen Einschränkungen betrachtet werden können. Beispielsweise muss man in unmittelbarer Nähe sein, um ein RFID-Tag anzugreifen. Ein ständiger Angriff, wie es beispielsweise bei Servern möglich wäre, ist ebenfalls kein realistisches Szenario. Die Basis des Verfahrens wird *pseudonym throttling* genannt. Die Idee dahinter ist, dass jedes Tag eine kleine Anzahl an Pseudonymen hat, mit denen es sich authentisiert. Jedes Pseudonym ist ein gültiger Authentisierungsschlüssel des

Tags. Die Pseudonymliste kann außerdem geupdated werden, sodass nicht immer die selben Pseudonyme verwendet werden müssen. Dies macht das Tag zunächst sicherer in Bezug auf die Privatsphäre. Angriffe wären jedoch durch einfaches Abhören eines Pseudonyms möglich. Um dies zu verhindern, muss sich der Leser zunächst beim Tag authentisieren, indem er einen Schlüssel $\beta_i$ zum Tag sendet. Dieser ist einzigartig zu einem Pseudonym $\alpha_i$. Jetzt authentisiert sich das Tag, indem er einen Schlüssel $\gamma_i$ sendet. Dieser ist ebenfalls einzigartig zu dem Pseudonym $\alpha_i$. Ein Pseudonym $\alpha_i$ hat also zwei Stellvertreter $\beta_i$ und $\gamma_i$, die zusammen eine Art Challenge-Response Protokoll ergeben. Nach einer erfolgreichen Authentisierung werden im Tag $\alpha_i$, $\beta_i$ und $\gamma_i$ geupdated.

Dieser Updateprozess könnte jedoch von einem Angreifer abgehört werden. Darum werden zusätzlich *One-Time-Pads*[5] eingesetzt. Das Senden von One-Time-Pads bietet jedoch keine Sicherheit, wenn dies über einen unsicheren Kanal geschieht. Darum werden die One-Time-Pads zusätzlich mit den tagintern Pads verrechnet:

Jedes Tag besitzt für einen Pseudonym-Stellvertreter-Tripel $k$ einen Vektor $\Delta_k = \{\delta_k^1, \delta_k^2, .., \delta_k^m\}$ aus One-Time-Pads. $\delta_k^1$ wird als *live* pad bezeichnet, da es $k$ updated ($k \leftarrow k \oplus \delta_k^1$).

Der Leser sendet nun ein Update $\tilde{\Delta}_k$ (= Vektor aus One-Time-Pads) an das Tag. Nun werden die One-Time-Pads in $\Delta_k$ mittels den empfangenen One-Time-Pads wie folgt verrechnet.

- Das *live pad* wird verworfen.
- Alle übrigen $\delta_k^i$ werden um eins nach unten geshiftet.
- $\delta_k^m$ wird mit Nullbytes aufgefüllt.
- Jetzt wird jedes Pad mit dem neuen Element des Vektors $\tilde{\Delta}_k$ XOR verrechnet ($\delta_k^i = \delta_k^i \oplus \tilde{\delta}_k^i$).

Da die $\delta_k^i$ unbekannt sind, ergibt sich mit der Verrechnung mit $\tilde{\delta}_k^i$ ein neues unbekanntes One-Time-Pad. Zu bedenken ist, dass $\delta_k^m$ nur mit Nullbytes verrechnet wird und daher eine Schwachstelle ist, dies wird aber kompensiert, indem es bei jedem Update nach unten geshiftet und neu verrechnet wird. Diese Methode bietet also einen relativ sicheren und einfach zu berechnenden Authentisierungsmechanismus, der für fast jede Art von RFID-Tags geeignet ist. Hardwarebasierte Verzögerung kann Tags dazu veranlassen, nur wenige Pseudonyme pro Zeitintervall preis zu geben, um das Auslesen aller Pseudonyme zu erschweren. Schutz der Privatsphäre ist ebenfalls gegeben. Eine verschlüsselte Kommunikation allerdings nicht.

Theoretisch ist das Verfahren nicht sicher, da das letzte One-Time-Pad nur mir Nullbytes verrechnet wird. Ein Angreifer, der alle Nachrichten im System abhört und auswertet, kann die verwendeten Codes berechnen und die Kommunikation entschlüsseln. Es wird daher angenommen, dass solche Szenarien nicht realistisch sind.

### D. $HB^+$

Hopper und Blum [13] haben ein Protokoll entwickelt, das die Basis für ein weiteres, für RFID-Tags mit geringer Rechenleistung geeignetes, Protokoll bietet. Der Ablauf ist folgender:

Es gibt zwei Partner $A$ und $B$ mit dem *shared secret* $x$. $B$ authentisiert $A$.

- $B$ sendet eine zufällige Zahl oder Bitstring (*Nonce*) $a$ an $A$.
- $A$ berechnet daraus ein Paritätsbit mittels dem Skalarprodukt $a \cdot x$ und sendet es an $B$.
- $B$ prüft, ob das Paritätsbit korrekt ist.

Ein Angreifer müsste hier jedoch nur das richtige Bit raten. Die Wahrscheinlichkeit, dass ein Dritter das richtige Paritätsbit errät, liegt also bei 50%. Um die Wahrscheinlichkeit zu verringern, wird der obige Vorgang $k$ mal wiederholt. Ein Angreifer könnte aber nach wenigen mitgehörten Runden mittels Gausselimination $x$ berechnen. Daher wird in jeder Runde das Ergebnis mit einer Wahrscheinlichkeit von $p \in (0, 1/2)$ verfälscht.

Die Authentisierung wird als erfolgreich angesehen, wenn mehr als die Hälfte aller Paritätsbits richtig sind. Das HB Protokoll sieht also wie folgt aus:

- Für jeder Runde $i$:
  - $B$ sendet eine zufällige Zahl oder Bitstring (*Nonce*) $a$ an $A$.
  - $A$ berechnet daraus ein Paritätsbit mittels dem Skalarprodukt $p = a \cdot x$.
  - $A$ verfälscht $p$ mit der Wahrscheinlichkeit $p \in (0, 1/2)$.
  - $A$ Sendet $p$ an $B$.
  - $B$ prüft, ob das Paritätsbit korrekt ist.
- $B$ prüft, ob mehr als die Hälfte aller Bits korrekt ist, falls ja, ist die Authentisierung erfolgreich.

Das Rauschen herauszufiltern, also $x$ aus mitgehörten Informationen zu berechnen, ist ein bekanntes mathematisches Problem[6], das NP schwer ist. Das HB Protokoll ist allerdings nicht gegen einen aktiven Angreifer geschützt. Sendet dieser immer die selbe Nonce $a$, wäre er in der Lage das Verfahren zu umgehen.

Daher wurde das HP Protokoll von Juels und Weis [12] zum $HB^+$ Protokoll erweitert.

Es hat zusätzlich einen *blinding* Faktor $b$, der eine vom Tag erzeugte weitere zufällige Nonce ist. Des weiteren gibt es ein weiteres *shared secret* $y$. A und B berechnen nun nicht mehr $p = a \cdot x$, sondern $p = (a \cdot x) \oplus (b \cdot y)$. Juels und Weis beweisen die Sicherheit ihres Verfahrens, indem sie das LPN Problem auf das HB Protokoll anwenden und auf das $HB^+$ Protokoll reduzieren.

Carrijo et al. [14] zeigten allerdings in ihrem Paper *A novel probabilistic passive attack on the protocols hb and* $hb^+$ einen Angriff auf beide Protokolle[7]. Daher ist das Protokoll als kryptografisch unsicher zu bewerten.

### E. AES

Feldhofer et al. [15] stellten bereits 2005 eine Technik vor, wie sich AES in *low-cost Tags* integrieren lässt. Da die

[5] One-Time-Pads sind kryptografisch sichere Verschlüsselungsverfahren, die keine komplexen Berechnungen benötigen. XOR Bitoperationen reichen für das Verfahren aus.

[6] Das LPN Problem: *Learning Parity in the Presence of Noise*

[7] Der Angriff geschieht mittels Gausselimination von Teilmatrizen. Der Grad des Rauschens muss dabei nicht bekannt sein.

Chips weniger rechenstark sind, müssen sie allerdings etwas modifiziert werden. Die Taktfrequenz muss beispielsweise erniedrigt werden, da die Berechnungen sonst zu viel Strom verbrauchen würden, um bestimmte RFID Spezifikationen noch zu erfüllen. Das Authentisierungverfahren funktioniert nach folgendem Schema:
Es gibt zwei Partner $A$ und $B$, die beide den selben privaten Schlüssel $K$ besitzen. $B$ authentisiert $A$.

- $B$ sendet eine zufällige Zahl oder Bitstring $R_b$ an $A$.
- $A$ verschlüsselt $R_b$ mit $K$ und sendet ihn an B. $R_a = AES_k(R_b)$
- $B$ verifiziert $A$, indem geprüft wird ob $AES_k^{-1}(R_a) = R_b$.

Ein fremdes Lesegerät hat jedoch die Möglichkeit den Tag zu orten, auch ohne in Besitz von $K$ zu sein. Sendet es immer das selbe $R_b$, wird auch immer das selbe $R_a$ zurückgesendet. Dies kann verhindert werden, indem A $R_a = AES_k(R_b + |+RANDOM)$ sendet. Gegenseitige Authentisierung ist mit diesem Protokoll ebenfalls möglich, dazu muss A nur einen Pseudozufallszahlengenerator besitzen. Die Authentisierung erfolgt dann für B analog. Viele Authentisierungsmechanismen basieren auf proprietären Protokollen, die durch ihren geheimen Algorithmus nicht gut auf ihre kryptografische Sicherheit überprüft werden können. Dieses Verfahren basiert auf einer 128 AES Verschlüsselung. Sie ist eine offizielle, lizenzfreie Verschlüsselung, die nach heutigem Stand der Technik als weitgehend sicher gilt. Eine Implementierung in EPC-Chips ist möglich, allerdings setzt dies eine leichte Modifikation der Hardware voraus.

### *F. Asymmetrische Authentisierung*

Asymmetrische Verfahren haben viele Vorteile. Mit ihnen ist es möglich, Schlüssel über unsichere Kanäle auszutauschen oder Nachrichten zu Signieren, um sie vor Fälschungen zu sichern und deren Ursprung zu verifizieren.

Radu-Ioan Paise und Serge Vaudenay [16] stellen ein Verfahren vor, das gegenseitige Authentisierung gewährleistet und dabei ein hohes Maß an Sicherheit und Privatsphäre bietet. Im Protokoll hat das System ein *Private-Public-Key* Schlüsselpaar $(K_s, K_p)$ und einen Masterkey $K_m$. Jedes Tag besitzt den öffentlichen Schlüssel $K_p$, eine eindeutige Identifizierungsnummer $Id$ und einen Authentisierungsstring $K$, wobei $K = F_{K_m}(ID)$

Die Anforderungen an das Kryptosystem sind allerdings sehr hoch, da es asymmetrische Verschlüsselung voraussetzt. Die Authentisierung zwischen einem Tag $T$ und dem System $S$ geschieht folgendermaßen:

- $S$ wählt eine zufällige Nonce $a$ und sendet sie an $T$.
- $T$ wählt eine zufällige Nonce $b$ und erstellt $c = Enc_{K_p}(ID|K|a|b)$ und sendet dies an $S$.
- $T$ entschlüsselt $c$. $ID|K|a'|b = Dec_{K_s}(c)$.
- $S$ prüft, ob $a = a'$ und $K = F_{K_m}(ID)$.
- Falls die Werte übereinstimmen, ist $T$ authentisiert.
- Nun sendet $S$ $b$ zurück an $T$.
- $T$ prüft, ob das empfangene $b' =$ gesendetes $b$.
- Falls ja, ist das System authentisiert.

| VERFAHREN | PRO | KONTRA |
|---|---|---|
| Baumstruktur | Für alle Tags geeignet | Schlüsselupdate aufwändig |
| Counter | Für alle Tags geeignet | Counter muss Systemweit bekannt sein |
| Pseudonym | Für alle Tags geeignet | Theoretisch knackbar |
| HB+ | Für alle Tags geeignet | Kryptografisch unsicher |
| AES | Offenes Protokoll, starke Verschlüsselung | Modifikation schwacher Chips notwendig. |
| Asymmetrisch | Sicher, Algorithmen frei wählbar | Nicht für alle Anwendungen geeignet |

Abbildung 3. Zusammenfassung der Algorithmen.

Die Id wird mit dem Masterkey $K_m$ verschlüsselt. Das Tag wird verifiziert, indem $K = F_{K_m}(ID)$ geprüft wird. Damit ist es möglich eine Tag-Id zu verifizieren, ohne dabei auf Datenbankeinträge zurückzugreifen.
$a = a'$ stellt sicher, dass die empfangene Nachricht als Antwort identifiziert werden kann.
$b = b'$ verifiziert das System, denn dieses kann den Wert nur korrekt auslesen, wenn es die Nachricht entschlüsseln kann. Die genauen kryptografischen Verfahren sind von Paise und Vaudenay nicht vorgeschrieben.

Für entsprechende Hardware bietet dieses Protokoll eine sichere Möglichkeit der Authentisierung. Für Szenarien, die einen hohen Sicherheitsstandard vorraussetzen und sichere Authentisierung benötigen, zum Beispiel in Hotels als Zimmerschlüssel, wäre ein Einsatz denkbar.

Schwachstellen sind allerdings auch hier gegeben. Die geheimen Schlüssel $K_s$ und der Masterkey dürfen nicht bekannt werden.

Abbildung 3 fasst die wesentlichen Aspekte der oben genannten Verfahren nochmal zusammen.

## V. FAZIT

Durch das relativ sichere Authentisierungsprotokoll von Paise und Vaudenay [16] lassen sich viele Probleme der Sicherheit und Privatsphäre eingrenzen. Es ist ein Verfahren, das mittels asymmetrischer Verschlüsselung eine sichere Authentisierung ermöglicht und die Privatsphäre des Nutzers schützt. Geht man nach *Moore's Law*, werden selbst kleinste RFID-Chips in der Lage sein, starke Verschlüsselungsverfahren und komplexe arithmetische Operationen ausführen zu können. Für viele Anwendungsszenarien ist es jedoch immer noch ungeeignet.

Authentisierungsverfahren mit *shared secret* sind eher für geschlossene Systeme geeignet und beispielsweise für Waren eines Supermarktes untauglich. Es lässt sich also nicht für alle Einsatzgebiete ein einziges Verfahren finden.

Schwachstellen sind häufig geheime, systemweite Schlüssel. Sobald diese bekannt werden, ist das komplette Verfahren unsicher. Ein Protokoll, dass absolute Sicherheit ohne Schwachstellen bietet, die Privatsphäre der Nutzer schützt und zudem gut skaliert, gibt es bisher nicht.

Sichere Chips werden nie vollständig den Markt beherrschen, solange einfachere, unsichere, und vor allem billigere Chips auf dem Markt sein werden. Weiterhin wird Schlüssel- und Passwortmanagement durch die vorgestellten Techniken selten gelöst. Dennoch gibt es Entwicklungen, die dem Nutzer den Umgang mit RFID-Chips erleichtern können. QR-Codes im Zusammenhang mit NFC-Chips in Mobiltelefonen beispielsweise, könnten diverse Passwortverwaltungs- und Sicherheitsaufgaben übernehmen.

Die Omnipräsenz und Vielfalt von RFID-Chips werden gerade bei der Privatsphäre viele Risiken aufkommen lassen, die uns möglicherweise momentan noch nicht bewusst sind. Ob diverse Vorkehrungen zum Schutz der Privatsphäre bei der *Generation Facebook* überhaupt noch verlangt werden, bleibt abzuwarten.

## LITERATUR


[1] A. Juels, "RFID security and privacy: a research survey," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 381– 394, Feb. 2006.

[2] E. Global, *Class 1 Generation 2 UHF Air Interface Protocol Standard Gen 2*, 2011 (accessed December 09, 2011). [Online]. Available: http://www.gs1.org/gsmp/kc/epcglobal/uhfc1g2

[3] N. Semiconductors, *Security of MIFARE Classic*, 2011 (accessed December 09, 2011). [Online]. Available: http://www.mifare.net/technology/security/mifare-classic/

[4] F. D. Garcia, G. Koning Gans, R. Muijrers, P. Rossum, R. Verdult, R. W. Schreur, and B. Jacobs, "Dismantling MIFARE classic," in *Computer Security - ESORICS 2008*, S. Jajodia and J. Lopez, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 5283, pp. 97–114. [Online]. Available: http://www.springerlink.com/content/t350426350037317/

[5] T. Kohno, "An interview with RFID security expert ari juels," *IEEE Pervasive Computing*, vol. 7, no. 1, pp. 10–11, Mar. 2008.

[6] D. V. Bailey, D. Boneh, E. Goh, and A. Juels, "Covert channels in privacy-preserving identification systems," in *Proceedings of the 14th ACM conference on Computer and communications security*, ser. CCS '07. New York, NY, USA: ACM, 2007, p. 297–306. [Online]. Available: http://doi.acm.org/10.1145/1315245.1315283

[7] M. R. Rieback, B. Crispo, and A. S. Tanenbaum, "RFID guardian: A Battery-Powered mobile device for RFID privacy management," in *Information Security and Privacy*, C. Boyd and J. M. González Nieto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3574, pp. 184–194. [Online]. Available: http://www.springerlink.com/content/3uq1qba2gcavmf20/

[8] A. Juels, R. L. Rivest, and M. Szydlo, "The blocker tag: selective blocking of RFID tags for consumer privacy," in *Proceedings of the 10th ACM conference on Computer and communications security*, ser. CCS '03. New York, NY, USA: ACM, 2003, p. 103–111. [Online]. Available: http://doi.acm.org/10.1145/948109.948126

[9] E. Biham, Y. Carmeli, and A. Shamir, "Bug attacks," in *Advances in Cryptology – CRYPTO 2008*, D. Wagner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 5157, pp. 221–240. [Online]. Available: http://www.springerlink.com/content/c728w668103776n1/

[10] D. Molnar and D. Wagner, "Privacy and security in library RFID: issues, practices, and architectures," in *Proceedings of the 11th ACM conference on Computer and communications security*, ser. CCS '04. New York, NY, USA: ACM, 2004, p. 210–219. [Online]. Available: http://doi.acm.org/10.1145/1030083.1030112

[11] A. Juels, "Minimalist cryptography for Low-Cost RFID tags (Extended abstract)," in *Security in Communication Networks*, C. Blundo and S. Cimato, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3352, pp. 149–164. [Online]. Available: http://www.springerlink.com/content/ny08l0b01ddn4jyf/

[12] A. Juels and S. A. Weis, "Authenticating pervasive devices with human protocols," in *Advances in Cryptology – CRYPTO 2005*, V. Shoup, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3621, pp. 293–308. [Online]. Available: http://www.springerlink.com/content/5443a0y0ej8c9bhp/

[13] N. J. Hopper and M. Blum, "Secure human identification protocols," in *Advances in Cryptology — ASIACRYPT 2001*, C. Boyd, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 2248, pp. 52–66. [Online]. Available: http://www.springerlink.com/content/4qrkua2x5adf29wu/

[14] J. Carrijo, "A novel probabilistic passive attack on the protocols hb and hb+." IEICE Transactions, 2008, vol. 3621. [Online]. Available: http://eprint.iacr.org/2008/231.pdf

[15] M. Feldhofer, S. Dominikus, and J. Wolkerstorfer, "Strong authentication for RFID systems using the AES algorithm," in *Cryptographic Hardware and Embedded Systems - CHES 2004*, M. Joye and J. Quisquater, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3156, pp. 357–370. [Online]. Available: http://www.springerlink.com/content/26tmfjfcju58upb2/

[16] R. Paise and S. Vaudenay, "Mutual authentication in RFID: security and privacy," in *Proceedings of the 2008 ACM symposium on Information, computer and communications security*, ser. ASIACCS '08. New York, NY, USA: ACM, 2008, p. 292–299. [Online]. Available: http://doi.acm.org/10.1145/1368310.1368352

# Driver Distraction in Automotive HMI

Marcel Reichersdoerfer
Ulm University
Ulm, Germany
Email: marcel.reichersdoerfer@uni-ulm.de

***Abstract*—Driver Distraction becomes more and more important to the whole world. Not only the amount of licensed cars on the road increases but every day at any place newer technologies are occuring and they are extending the functionalites of a car. Furthermore many people are bringing some system into the car for using it while driving for example mobile phones. But who is caring for observation of safety aspects so that driving doesn't become more dangerous than now. This paper shows different systems speech recognition or handwriting recognition in their usage and whether they are profitable for the driver because of distraction problem. Also certain standards and guidelines are existing to reduce the probability to produce a distractible system. In the end there is a research about a improved head-up display.**

## I. Introduction

In these days Driver Distraction is a very important theme to automotive industry. Over the whole world there were and are still both little and big changes in the environment. At Figure 1 you can see that the number of cars on the road have increased enormous by more than 20 million for the last fifty years. Additionally the lengths of the route people are driving raised up to above 400 billion kilometers.

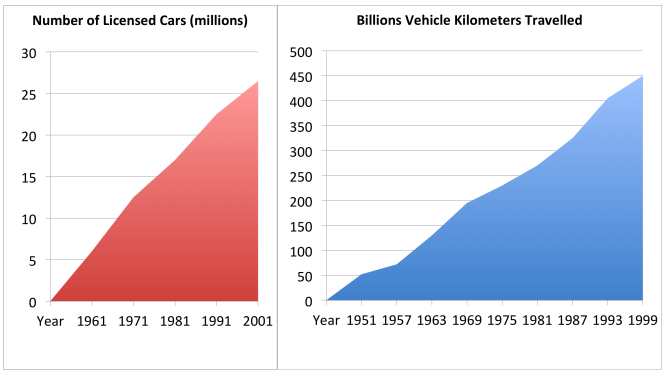


Figure 1. Statistics [1]

This also led to the development of more intelligent innovations and newer technologies. It should be added that new systems attract people to stay up to date what belongs to features of their cars. For example with the aid of navigation systems or Personal Digital Assistants (*PDAs*) the driver has a map to look at and even the route to drive will be calculated if the driver enters the start and end point. So ordinary maps become needless but at the same time the inattention to the road traffic raises dramatically.

Also the use of smartphones that are the latest generation in the mobile phone industry got a very high attention. In the past calling other persons and sending text messages (*sms*) was the only thing mobile phones could do. Nowadays that are the minor used functionalities because even writing e-mails, surfing in the internet and the use of applications installed on the smartphone and depending on the internet for example Facebook does not pose a problem anymore.

The mainly reason for driver distraction belongs to the interaction between the driver and the used device in the car, in short Human-Machine-Interaction or Human-Machine-Interface (*HMI*). To interact with a machine it needs an interface which can be operated through the driver. There are different types of interfaces possible, most of the time it is a haptic or visual one sometimes also based on speech or gesticulation whereby the latter have more effect to distraction. In Figure 2 you can see a interface of haptical type, the control gear too.


Figure 2. Haptic Interface [2]

But for all these things it should not be forgotten that driving is the primary task while sitting in a car. When moving the driver's hands are on the steering wheel so the distraction level becomes higher and higher for each new function that cars have especially if it requires a look at the device for haptic interaction. So this paper shows rules and guidelines reducing the danger of in-vehicle information and communication systems. Furthermore a detailed description of automotive parts like visual interfaces relating to distraction will be discussed regarding how distraction occurs.

## II. State of Distraction

In recent years there were many changes in the automotive industry because of new developed Technologies. Electronic and telematic devices became multi functional, especially devices became increasingly portable and don't belong any longer to phone lines or vehicles. Not only original and aftermarket automobile equipment is in Production but also electronic devices for entertainment are manufactured. This causes disagreements in the automobile industry regulatory structure. Additionally new products are developed, introduced and modified at a great pace [3]. So the distraction level raises to an enormous value.

At the primary task which is driving the driver looks onto the street and is alert. So further tasks have to be time-shared. Thus the driver has to switch back and forth between looking and doing a secondary task. Therefore the blind driving time range from 0.6 to 1.6 seconds, what means an average of 1.2 seconds. If you gross it up to a trip with 30 mph, a length of 50 feet blind driving is reached [3].

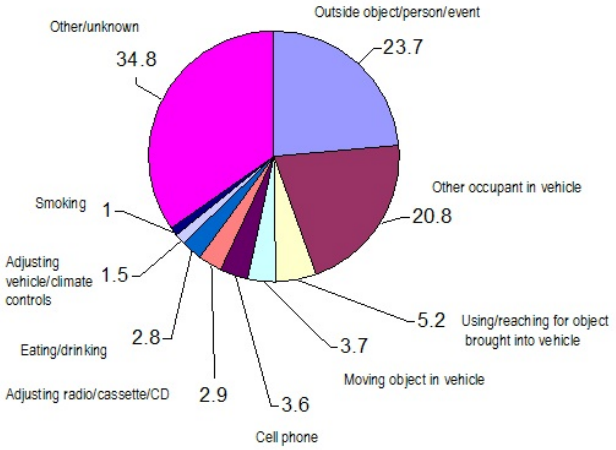


Figure 3. Distraction statistic [4]

In Figure 3 a statistic about reasons for crashes are given. 1 percent of the people were distracted by smoking while driving, nearly 2 percent justified their controls. Almost 3 percent were eating or drinking when the crash happens and further barely 3 percent adjusted their radio, cassette or CD. People distracted through their mobile phone amounts to 4 percent and comes even with them who were confused by an object in their vehicle. More than 5 percent wanted to get an object they brought into the car, nearly 21 percent were distracted by other passengers, almost 24 percent lost attention through an object, event or person outside the car and the rest were other unknown things which distracted from the road.

### A. Navigation Systems and Music Players

Generally distracting effects of operating entertainment systems were considered as harmless. But studies have shown that only turning radio on or listening to the radio while driving can distract drivers[5]. Even operating with a CD Player while driving is more dangerous than dialing a mobile phone or eating. Also navigation systems are distractive because the driver has to look onto the screen when entering a destination or only viewing for direction[5].

### B. Mobile Phones

In a survey 81 percent admit talking with passengers during driving with the car[3]. Just recently there are Smartphones and meanwhile a high amount of peoples are using them. If you compare both situations you can clearly say that a few of them are calling other persons while driving although it is against the law. The issue is that the main attention goes to the person on the other end of the line. Furthermore if they don't own a hands-free the point of danger is raising. Additionally only one hand can touch the steering wheel so the chance of timely reaction in an accident is extreme low. Even the opportunity to drive past a car crash happening right on the front of your car is very low. But since smartphones exists there are also applications for your phone to install. So if the driver of a car hears a signal from a special application he loses his concentration to the road traffic.

### C. Speech Recognition

At the moment there are two limitations to speech-based systems[3]. Think of driving across the streets and you are saying to your system to play a special song you want to hear. If you are a little bit faster in transit maybe there are wind and road noises, within a city or town there are traffic noises. So the first limitation is the minimized chance that the system understands the commands of the driver. Only if the radio is turned on it will be complicated for the system to recognize the signal correctly. As a result the driver pays more and more attention to the system and at the same time the distraction level increases on and on.

The other theme is a user related limitation. Certainly there are drivers with any disabilities preventing from the use of the system. Some people are driving as well as never alone. So they don't like to interact with a speech-based system because they are shamed of talking to a technical device due to bad feelings generated while another person is listening to this conversation between human and machine. Another crowd perhaps complains about learning too many commands to use the system, especially when they are under high workload conditions.

So the ideal solution would be a system that don't exacts specific commands but understands the normal language often referred to Natural Language Processing.

### D. Handwriting Recognition

Handwriting recognition systems use a physical touch pad that recognizes alphanumeric text data entry[3]. It is easy for the driver to draw a character onto the pad without loosing the view to the street. But for recognition of the system that the input is recognized the driver has to look onto an


Figure 4. Distraction statistic [6]

interface again to check it. Because of left hand and right hand manufactured vehicles for the market researches are necessary to find out if the use of the steering wheel location or touch pads at the center control and door side is the better solution. In Figure 4 the handwriting recognition on a mobile phone is shown. The user writes a word onto the screen and the system recognizes the word.

Despite there are problems at the detection of digits, e.g. the difference between '1' and 'I' or '5' and 'S'[3]. So many peoples wouldn't accept a system like this. Also feedback from the system is a great thing. Only audible responds instead of using interfaces would be necessary to reduce the distraction. So the best solution would probably be that the system gives feedback by spelling the entered word or phrase. This leads to next issue, the input acknowledgement. What would be the best idea to confirm the input consisting of several characters? Either the user draws a arrow symbol like on computer keyboards or they can find an extra button to confirm.

Even if there are always a few issues connected to handwriting recognition it possesses the opportunity to switch to only manual-based systems so that the necessity of visual interfaces disappears.

### E. Gesture Recognition

There are two opportunities to gather input from the user[3]. The first one is trying to get by a three-dimensional model of the users hand the needed information. The other option are image based methods where flesh tones from the background images were segmented to find hands and then try and extract features such as fingertips, hand edges or gross hand geometry for use in gesture recognition.

All automotive Static and most Dynamic, non-contact hand gesture research found, uses vision based technologies, however, it is well documented that some technical challenges remain to be resolved, namely coping with dynamic backgrounds, variable lighting conditions and response times. In the long term, it is expected that the vision based approach to automotive hand gesture recognition will be adopted by many automotive OEMs since it is likely the above mentioned technical difficulties will be overcome[7]. Furthermore it is a problem that the range of hand gesture within a car is limited.

So in summary you can see that gesture recognition would be a option but not really the best one because it can endanger driving.

## III. Standards and Guidelines for HMI

To guarantee the safety of driving with supportive systems without producing the distraction of the driver different guidelines like the European Statement of Principles and the Alliance of American Manufacturing were developed. Both of them are very similar to each other.

### A. European Statement of Principles

In 2006 the European Statement of Principles on Human Machine Interfaces for in-vehicle information and communication systems was formed. Their usage require both technical knowledge of the products and the opportunity to implement them in system design. Also they do not prevent innovation of the industry. The principles are not a substitute for any current regulations and standards, which should always be taken into consideration. These principles can be reinforced by national legislation or by individual companies. These principles constitute the minimum set of requirements to be applied [8].

The main focus of the principles belongs to in-vehicle information and communication systems when drivers are using vehicles while moving for example navigation systems or mobile phones. Because of insufficient research results and scientific proof the scope of them excludes stabilization systems (e.g. ABS or ESP), voice controlled systems or systems offering information, warnings or support through the driver's active manipulation. As well head-up displays and aspects separate from HMI as electrical characteristics, material properties and legal aspects not related to safe use. As long as no Differences are made the principles concern only system use by the driver while driving. The responsibility belongs to the provider of the system. If the provider falls apart into several groups or organizations also the responsibility will be distributed.

But the general field of attention in this subsection applies to the principles, so let's have a look at them [8].

*1) Overall Design Principles:* The system has to assist the driver and prevent of producing dangerous manner. In this way road safety can be guaranteed because designers are concentrated to the appearance of installation, information presentation or interface. The Decision to interact with their system has to be the driver's task. Also the moment and the way of use are on the driver to do without loosing attention to road traffic. The importance of avoiding distraction caused by visual entertainment because of the safety of the driver has to be considered. Furthermore information given by the

system should not provoke the driver. The meaning of this is to discourage driver for inconsiderate action and thereby reducing accidents. Additionally a comparison between the old systems and the developed one is essential. The difference should be minimized so that drivers aren't confused by entirely unusual design.

*2) Installation Principles:* For installation of systems into cars you have to notice established regulations, standards and manufacturers instructions. So this includes that the system cannot be used in another way as manufacturers ordered and prevent of false function ergo safety aspects. Moreover the vision of the driver mustn't be interfered through parts of the installing system (except head-up displays). Thereby vision means the windscreen, side views and rear view. If there is the chance to move out of the sight through manipulation by the driver the manufacturer have to give information about how to do. Also the view of displays and controls needed for the primary task of driving mustn't be obstructed by the system for example accelerator, brake or steering wheel. As more important a system is for fulfilling the driving task as closer it has to be in the line of the sight of the driver because of Danger of having too much eyes-off-the-road time. In addition no reflections should occur through the position of the system within the car.

*3) Information Presentation Principles:* The presentation of information interesting for the driver has to be as short as possible so that the moment of looking onto the screen doesn't influence driving. Besides the use of international/national legibility, audibility, icons, symbols and so on is very important. The given information has to be correct and showed for a suitable time period. At the same time priorities about the safety notability of information is indispensable. Sounds of the system that can't be changed by driver mustn't drown the voice of warnings both inside and outside of the car.

*4) Interfaces with Displays and Controls:* While interacting with the system the driver have always to be up to possess one of his hands onto the steering wheel. If the system presents manual-visual interface sequences it should be hold in a briefly and non-interruptible form as possible and avoid skipping quick procedures. At the same time the driver is said to have the opportunity to continue a process at a consistent or directly on the interrupted point. Also the driver is believed to decide on his own when he wants to give an answer to system demands instead of time-based reply. The design of system controls may not affect the use of primary driving controls. Sound intensity of auditory information has to be adjustable by driver if it exists the probability of deviation. As well the driver shall recognize obvious and in a curt time per system response his entered action to the system. Beyond the system has to enclose a mode for stopping the offer of dynamic visual information that doesn't discharge safety issues.

*5) System Behavior Principles:* Visual information that is not important for driving task but distract notably the driver have to be deactivated or not invisible for the driver during car is moving. In addition the system's manner mustn't impair other displays or controls necessary for primary driving task or road safety. The chance of Interaction with system functions not accomplishing their purpose while driving mustn't exist if the vehicle is in motion, as alternative distinct warnings of no appropriate use are allowed. All the time the driver has to be in the position to see the current status of the car and each system failures probably interfering the safety of driving.

*6) Information about the System:* At all a manual including ordinary use, significant facets of installation and service or caring for the system got to exist. Besides this instruction has to be correct and easy to understand. Additionally the language and the form of the manual should be well shaped and consequential suitable for the group of drivers. Because of the instructions the driver must know which functions of the system are designed for use while moving with car and which ones not. Moreover product information got to explain in an exactly way how system functionalities are doing. Furthermore the product information has to mention if specific abilities are necessary for the use of the system as manufacturers specify or if some users mustn't use it. Also representation of system use should avoid visionary expectations on user side and discourage unsafe use.

In summary there are six different types of rules. The *Overall Design Principles* determine the design of the system in a manner not to influence the driver in a negative way. By the *Installation Principles* the position of the system will be defined. The *Information Presentation Principles* set what the driver is allowed to see. Also about the *Interaction with Displays and Controls* certain regulations are existing. Especially the status of the system has to be obvious ensured by the *System Behavior Principles*. In the end a documentation for help or information for every system is necessary given through the *Information about the System* Principles.

### *B. Alliance of American Manufacturing*

In 2001 the National Highway Traffic Safety Administration got a comprehensive draft of guidelines for design, installation and use of future telematic devices by the Alliance of American Manufacturing. With the aid of specialists these guidelines became more comprehensive regarding performance criteria and verification procedures. The particular emphasis of this design principles belongs to light vehicles and their Intention applies to both original equipment manufacturers as well as the aftermarket.

These principles concern safety aspects of Human Machine Interfaces for "new" information and communication technology and devices with visual and manual/visual interfaces, features and functions designed to be used by a driver while driving (vehicle speed $\sqcup$ 5 mph) and under "routine driving conditions". The scope of them doesn't bear on head-up displays, voice-activated devices, haptic displays and cues, purely cognitive distraction (e.g. conversation) and driver assistance systems. Also there are no differences

between them and the European Statement of Principles but they differ in their definition.

## IV. Research

The dynamic active display (DAD) is a large-area windshield display for warning the driver in critical situations by an alert [7]. Thereby a laser is used to place dynamic visual icons to the windshield. Also the option to change the intensity and position on the screen exists. At the moment heads-up displays are designed to present information to the driver. But through a DAD the driver doesn't have to look down to see the information but can spend more time looking at the road.

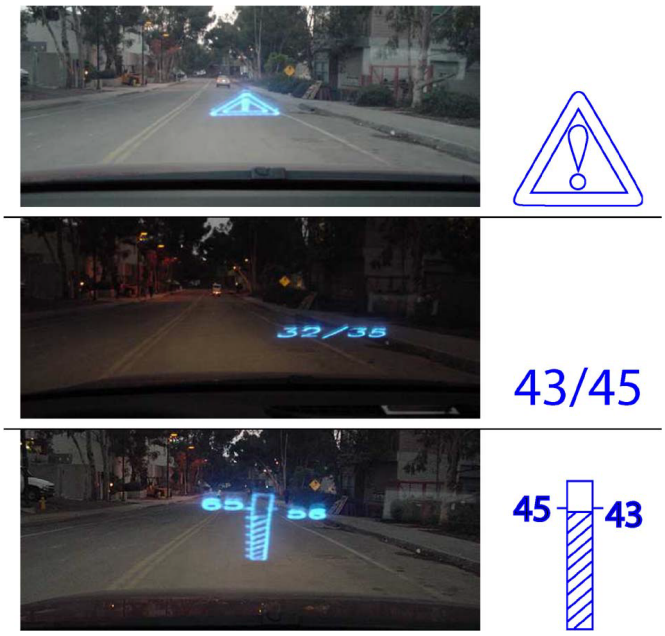


Figure 5. Three strategies [7]

By dynamic displays the driver possess for all the time about an overview of the state of the vehicles and the environment [7]. Also navigation systems and dashboard warning lights are dynamic showed. So an active display shows information and monitors the state of the driver.

### A. Experiments

At least four strategies of speed alerts were tested. Therefore each driver had to drive the same route and was familiar with the area. Regarding measures the amount of time the driver spent above the speed limit, the ratio of time spent observing the alert or dash or the road in general and the distribution of speeds measured for roads with various speed limits played an important role [7]. The following strategies are shown in Figure 5.

- No display - No DAD alert is given.
- Warning - A triangular exclamation point warning sign appears and bounces as soon as the driver exceeds speed limit.
- Numbers - A textual alert constantly shows the driver's current speed and the road speed limit (e.g., 43/45). The text representing the driver's speed zooms in and out if the driver is above the speed limit.
- Graphic - A graphical alert constantly shows a vertical status bar with the driver's speed and the speed limit clearly marked. The entire graphic bounces if the driver is above the speed limit.

### B. Results

To represent how fast the effects of different warnings affected the drivers the average amount of time the driver spent over the speed limit before returning to under the limit were measured. For all test subjects, the caution symbol from the second condition caused a drop of 2.24 s in the average time to slow back down [7]. Then the values were compared with condition 1. The second display condition caused a drop of 38 percent in the time to slow back down. The other conditions number 3 and 4 were effective but not as much as the warning condition 2.

At next there were measures for the distraction. The result shows three fields of view, the windscreen, the DAD and the dashboard. Condition 3 caused with 37 percent of the normal time looking down the best result but at the same time increased the time spent with looking forward by 10 percent. The warning was the most helpful condition because it caught the attention of the driver better than the others. Additionally it was able to tell the driver immediate if he was driving above the speed limit without causing them to switch their view away from the windscreen.

As a result a DAD is very helpful and can reduce the distraction level by alerting drivers without changing their look.

## V. Conclusion

Because of the increasing number of cars on the road and the raising use of information systems in the vehicle driver distraction becomes a very important theme around the world. So at first every driver has to fullfill the primary task of driving. Many people take their own devices from home into their car think of mobile phones which would be the most common object. But through calling or using applications on a smartphone their grade of attention is falling down. Also operating with a CD-player is dangerous. Even operating entertainment systems were considered as no problem for distraction although studies tells the opposite. So the chance of losing attention of the road traffic is enormous.

Other systems like handwriting recognition or gesture recognition would be a great idea but they also have their negative aspects. First one because it possesses the opportunity to switch to only manual-based system what means that there is no need for a visual interface any more. The other one is not really the best too because the driver has to move with his hand to inform the system what to do and this disagrees the primary task of driving. So the best one would be the third recognition system namely speech. The driver can look onto

the street while using the system so the distraction would be a minimum. But at the moment no natural language is possible thereby the system is not perfect.

Furthermore standards and guidelines are a very important regulation for safety. Because of principles for installation, design, information presentation and so on a minimized distraction possibility can be guaranteed.

Despite explorer are searching for even better solutions to reduce the distraction of drivers when using systems while driving and to perfect available systems to the most minimal distraction level. The dynamic active display could be one of this solutions.

So surely can be said that driver distraction will be a inescapable point for the next few years.

## REFERENCES

[1] "Statistics," Accessed: Dec 8th, 2011. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1423417

[2] B. T. Ansgar Meroth, *Infotainmentsysteme im Kraftfahrzeug*. Vieweg+Teubner Verlag, 2008.

[3] "A REVIEW OF AUTOMOTIVE HUMAN MACHINE INTERFACE TECHNOLOGIES AND TECHNIQUES TO REDUCE DRIVER DISTRACTION." [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4399937

[4] "Distraction Statistic," Accessed: Dec 16th, 2011. [Online]. Available: http://safety.transportation.org/htmlguides/DDD/Exhibits/exhibit_III-2.gif

[5] "THE ROAD SAFETY MONITOR 2010 distracted driving," Accessed: Dec 14th, 2011. [Online]. Available: http://www.tirf.ca/publications/PDF_publications/RSM_2010_Distracted_Driving_Eng_3.pdf

[6] "Handwriting Recognition," Accessed: Dec 13th, 2011. [Online]. Available: http://cdn.macrumors.com/article/2008/05/06/153937-500_DIOTEK.jpg

[7] "A Novel Active Heads-Up Display for Driver Assistance." [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4581392

[8] "COMMISSION RECOMMENDATION," Accessed: Dec 5th, 2011. [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:032:0200:0241:EN:PDF

# Gamification: State of the Art Definition and Utilization

Fabian Groh
Ulm University
Institute of Media Informatics
Email: fabian.groh@uni-ulm.de

***Abstract*—This paper analyzes a definition of gamification as "the use of game design elements in non-game contexts". In this context, gamification can be set in relation to and differentiated from exisiting research. Due to the fact that this definition allows various game design elements, these are related to the self-determination theory. Consequently, there is a discussion about what gamified applications are able to offer and what problems they indicate.**

## I. Introduction

Current developments show the success of gamified applications, like the location-based service Foursquare[1] or the protein folding game foldit[2], where gamers solve the crystal structure of a retroviral protease which is involved in reproduction of HIV [1]. Analyzing these examples, a phenomenon can be observed: user activity and retention are improved by combining game design elements with non-game context. This phenomenon is called "gamification" in most cases.

While this term was first introduced in the early 2000s, it did not get huge attention until different industry players as well as conferences put it on the map for a wider audience in the second half of 2010. Thereupon, plenty gamified applications were developed across different domains such as productivity, finance, health, education, sustainability as well as news and entertainment media. Of course this also caught the attention of digital marketeers and interaction designers to improve their products. For example several vendors offer gamification as a software service layer of rewards and reputation systems like levels and badges. As we can see, the commercial aspect of gamification has a wider audience and it mainly differs between two concepts. The first point is the increase of acceptance and ubiquity of games in everyday life [2, 3, 4] and the second is to improve non-game products by increasing the motivation of users to engage with them on the same level of unique intensity and duration which is known from game design elements [5, 6]. Additionally, this also offers new possibilities and data sources for human-computer interaction.

Although, the phenomena of gamification is recently gaining the scientific attention [7, 8] [9], only few works provide a definition of gamification. Therefore, we firstly have a deeper look into the work of Deterding at. al. [10, 11]. They survey and situate current uses of gamification within existing research and explain as well as take apart their definition of gamification as "the use of game design elements in non-game contexts". Thereafter, we analyze the criteria which should be fulfilled by an gamified application. The last part then discusses the chances as well as the threads of gamification.

[1]https://foursquare.com/
[2]http://fold.it

## II. Definition

This part mainly follows the work of Deterding et. al. [11, 10], who define gamification as "the use of game design elements in non-game contexts". If gamification has to be identified as an academic term, it is necessary to demarcate this phenomenon from previous research as well as embed it into existing fields. Therefore and in order to better understand gamification, we first need to clearly define the related terms of this definition.

### A. Game

First, we have to distinguish between game and play. Whereas the word "play" is used for games as well as for toys, there is a difference. Therefore, we follow Caillois' concept, which describes paidia and ludus as two poles of play activities [12]. While paidia (playing) describes free-form, expressive, improvisational behaviors and meanings, ludus (gaming), on the other side, characterizes rule-based playing with determined goals.

In matter of gamification we are talking about the design elements of gaming (i.e. ludus) since critiques from the academic as well as the industrial point of view have indicated the focus of gamification is almost exclusively on ludus with just a little space for paidia [13]. Consequently, the term "gamfulness" is taken over from McGonigal [7]. She has introduced "gamefulness" as a counterpart for "playfulness". While, "gamefulness" describes the experiential and behavioral qualities of gaming (ludus), "playfulness" describes the same for playing (paidia). Hence, in the HCI context gamification is distinguished from playfulness, playful interaction and playful design [14, 15]. Nevertheless, in practice gamified applications can also encourage playful behaviors and mindsets.

Additionally, gamification should not be limited to digital technologies. Whereas the majority of gamified applications is digital, this constraint is not required.

### B. Element

In contrast to "serious games", which are full-fledged games for non-entertainment purpose, gamified applications just use elements of games. On "Serious games" the main goals are solving problem to train, investigate or advertise the users [14]. Examples are "America's Army"[3] for the purpose of train and advertise recruits, and "Microsoft Flight Simulator"[4] which was created as a simulator for civil aviation. However, this distinction may be very personal, social and subjective and depends on your perceptions and enactments. Hence, it is not clear if a group of people 'play' or 'use' an application like for example Foursquare. In their work, Deterding et. al. also found increasing acknowledgment that a definition of game must go beyond properties of the game artifact to include situated and socially constructed meanings to handle applications like Foursquare where the aspects of 'play' and 'use' are mingled [11]. Therefore, we have to (a) look for technical as well as social elements of games and (b) technical elements should be more as affording gameful interpretations and enactments rather than being gameful.

The next step is to clarify what is the set of 'game elements'. For example, take some of the "Ten Ingredients of Great Games" by Reves and Red [8] like avatars, time constraints, feedback, ranks or levels. Whereas if one of these elements are common in one type of game, it could be very unlikely to be found in another kind of game, e.g. avatars and levels are usual in adventure games but there are rare in real-time strategy games. However, non of such elements would automatically identify an application to be gamified. Additionally, all of them can already be found outside of games. In matter of define gamification there are two opposite poles to limit the set of game elements. On the one hand any element found in any game (liberal), and on the other hand elements that are unique to games (constrained). However, this would be either boundless or empty. Thus, Deterding et. al. suggest to limit gamification to the description of elements that are characteristic to games, but this is a heuristic and there is still much room to discuss [11].

### C. Design

As already mentioned, elements of games can be found outside games. If we are looking more into the technical aspects, there is for example a long tradition in HCI to use game controllers and other hardware for non-game purposes, e.g. XBMX[5] a media center for the video game console xbox, as well as using graphic engines and three dimensional environments for simulations. Hence, the term of game *design* is used instead of game-based technologies or practices.

[3]http://www.americasarmy.com/

[4]http://www.microsoft.com/games/flightsimulatorx/

[5]http://xbmc.org

Table I
LEVELS OF GAME DESIGN ELEMENTS.

| Level | Description | Example |
|---|---|---|
| *Game interface design patterns* | Common, successful interaction design components and design solutions for a known problem in a context, including prototypical implementations | Badge, leaderboard, level |
| *Game design patterns and mechanics* | Commonly reoccurring parts of the design of a game that concern gameplay | Time constraint, limited resources, turns |
| *Game design principles and heuristics* | Evaluative guidelines to approach a design problem or analyze a given design solution | Enduring play, clear goals, variety of game styles |
| *Game models* | Conceptual models of the components of games or game experience | challenge, fantasy, curiosity; game design atoms; |
| *Game design methods* | Game design-specific practices and processes | Playtesting, playcentric design, value conscious game design |

Deterding et. al. have analyzed the description of game design elements from previous researches and out of that they defined their own five levels of abstraction, which are shown in Table I from concrete to abstract [11]. They suggest to include all these levels in the definition of gamification.

### D. Non-game contexts

Game design elements are normally created for the purpose of entertainment. However, for the case of gamification these elements are now used to enhance other applications, similar to serious games [11].

Nonetheless, Deterding et. al. suggest to not limit the definition of non-game context any further for any specific usage, intentions, contexts, or media, because (a) there is no clear advantage of such a restriction and (b) serious games have also developed themselves into all kind of contexts, although there was a specific purpose (e.g. learning) at the beginning. Thus, Sawyer's taxonomy of serious games [16], which provides training games, newsgames, or health games as potential subcategories can also be adopted for potential subcategories of gamification: training gamification, news gamification, health gamification, and other application areas.

Furthermore, this definition for gamification also excludes the presumed gamification of games [17] with a meta-game platform like an achievement system [18], because this would simply be an extension of a game and consequently a part of the game design and not gamification.

### E. Situating "Gamification"

The definition of gamification from Deterding et. al. can be summarized the following way. Gamification is:

- the use of (in contrast to extension)
- game (in contrast to play or playfulness)
- design (in contrast to game-based technology)

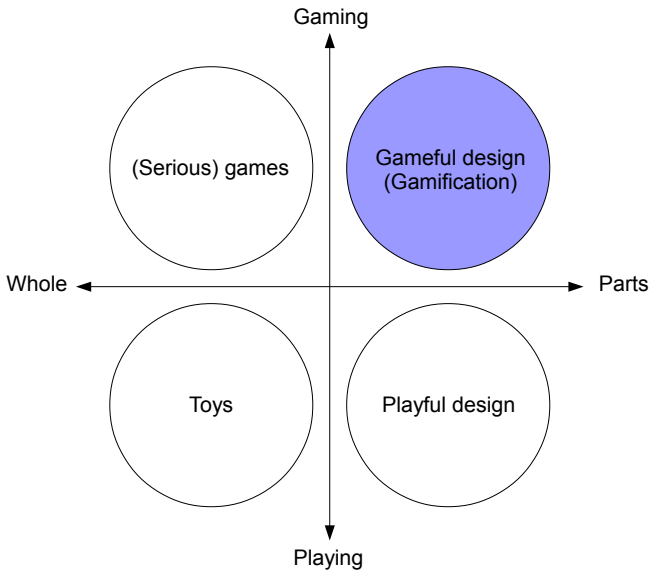


Figure 1. Separating the term of gamification from serious games on the parts/whole axis as well as from playful design along the gaming/playing axis [10].

- elements (in contrast to full-fledged games)
- in non-game context (regardless of specific usage intentions, contexts, or media of implementation)

As shown in Figure 1, the definition from Deterding et. al. separates gamification from serious games and the design of playful interaction on the two dimension of playing/gaming and parts/whole. While, the parts/whole-axis separates games and respectively serious games from gamification, the playing/gaming-axis differentiates between gamification (gameful design) and playful design as well as toys if both axes are used.

If a broader scheme is used to place gamification, it can be found in the socio-cultural trend of ludification. This is illustrated in the Figure 2.

There is one alternative academic approach for defining gamification from Huotari and Hamari [19]. They use a service-marketing point of view to define it as a "service packing where a core service is enhanced by a rules-based service system that provides feedback and interaction mechanism to the user with an aim to facilitate and support the users' overall value creation". Deterding at. al. indicate that, if the definition from Huotari and Hamari would be taken, even an touchpad for ordering snacks in a cinema would be a gamified application. Therefore, their own definition would improve this weakness.

## III. Principles for Gamifying Applications

This chapter is inspired by two presentations from the "GoogleTechTalks" program[6]. The first presentation "The Pleasure Revolution: Why Games Will Lead the Way" is presented by Jesse Schell [20] and the second one

[6]research.google.com

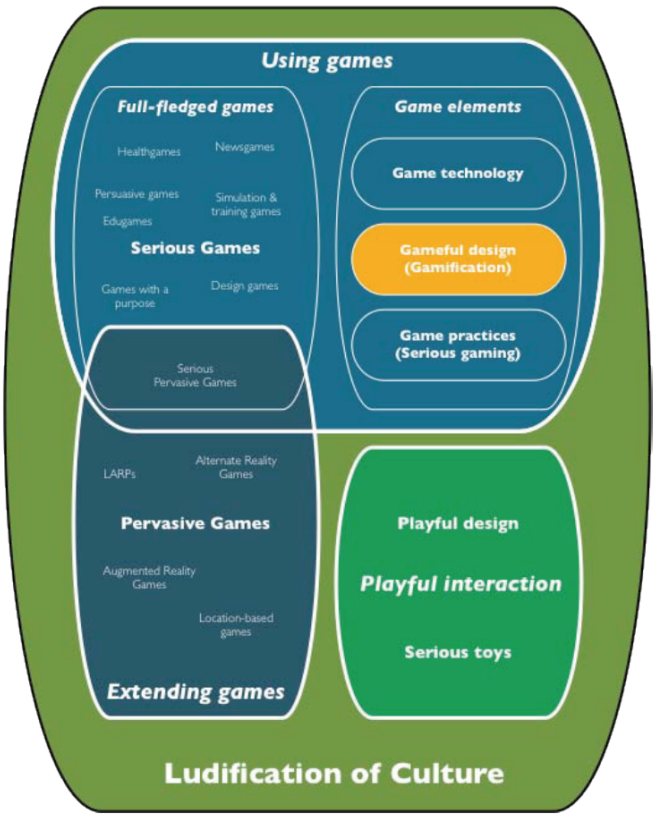


Figure 2. Placing the term gamification into the broader schema of ludification to get the relations. Graphic from Deterding [10].

"Meaningful Play: Getting Gamification Right" by Sebastian Deterding [21]. These speeches present some principles for the elements of game design for a good gamified application. Unfortunately, there is no published print version about the presentations until now due to the novelty of the field.

First of all, Schell and Deterding both criticize the state of the art of gamification, as a blueprint for putting point, badges, and leaderboards on everything. However, pleasure is not additive and rewards can backfire. This is shown by Alfie Kohn in his book "Punished by rewards" [22]. One of the studies demonstrates that children will draw more pictures, but in lesser quality, if they are paid for drawing pictures. However, the more important fact is that the children did not like drawing pictures as much as before, after they are stopped being paid. This effect is know as "Overjustification" and verified by Lepper et. al., where intrinsic motivation is shifted towards the extrinsic incentives [23].

Therefore, Schell and Deterding introduce three important principles, which are adopted from the "self-determination theory" from Deci and Ryan [24]. This theory describes three innate needs for intrinsic motivation:

- **Relatedness**: The universal need to interact and be connected with others.
- **Competence**: The universal need to be effective and master a problem in a given environment.
- **Autonomy**: The universal need to control one's own life.

The connection between the "self-determination theory" and games is shown in "Glued to Games" by Rigby and Ryan [25], where they also demonstrate that games are doing well in satisfy this needs.
Thus the next three paragraphs explain these points in detail.

### A. Relatedness

After its huge success, more and more foursquare users are leaving the service, because they have realized that there is no real benefit beyond the rewards. Comparing this to a site like "stackoverflow"[7], a question and answer platform mainly for the purpose of programming. Even after removing all points and badges from the platform, there would still be meaningful content. For that reason, it is important to catch the user's *personal goals*, or more general, customizable goals, which are connected to any interest or passion of the user that he already has in his everyday life. For example, the mini game foldit, mentioned in the introduction, primarily engages a lot of people with a scientific background.
This directly leads to the next point. In order for status and reputation (levels, badges, leaderboards) to work, it is important to connect the user to a *meaningful community* with the same interests. An achievement is made to show it your friends with the same interests. If there is nobody whom you can show it to, your achievement will be nothing special.
A common method of video games to reach more relatedness is a *meaningful story*. In most cases this means that the player has to save mankind from something. This approach can also be used in real-life applications by wrapping a story around the activity, which is further supported with suitable visuals.
However, in the context of relatedness it is important to *beware of social context meaning*. While something is clear on different platforms or within a certain group, it can be confusing for people outside this area.

The principle of relatedness can be summarized as:

- Connect to personal goals
- Connect to a meaningful community of interest
- Create a meaningful story
- Beware of social context meanings

### B. Competence

The need of effectively solving a problem depends on problem itself. Therefore, Schell introduce an idea of a button, where somebody who needs to clean dishes could simply push the button and all the work would be done. A great number of people would like to have such a tool. However, if this button would also exist to win a game by simply pushing it once, this game would probably have not much success. The "Progress Wars"[8] from Jakob Skjerning is such a game, where the only task is to click a button, which is not really engaging (Figure 3). Thus, the set of tasks can be divided into the categories of "Have to do" and "Want to do". Where the category of "Have to do" includes things like duty, work, slavery, and even efficiency, whereas the "Want to do" involve fun, play, freedom, and especially pleasure. While the first category should be as simple as possible, the second category should not.


Figure 3. Progresswars: A parody of game design, where you get rewards and levels by simply clicking one button.

Furthermore, the game designer Raph Koster said: "Fun from games arises out of mastery. It arises out of comprehension. It is the act of solving puzzles that makes games fun. With games, learning is the drug." [26]. This means that the fun in video games is to develop an ability to master the next step in the game. While, solving mathematical equations in the context of games is fun for a lot of people, they would simultaneously disagree to denote the same task as fun in the context of school, where most of the tasks seem to be boring and repeating. Additionally, in the context of games, people often even do not notice or realize that they are doing math.
Thus, it is important to confront the user with *interesting challenges*. This can be reached by combining well-defined goals and rules.
Moreover, it is crucial to present these goals in a *clear and visual* way as well as *structure* them well. That means the goals are taken apart into smaller chunks in such a way that there are always small and doable tasks. The *scaffolding* for those challenges should increase the difficulty for reaching the next level. Additionally, it should match the "Flow theory" from Csikszentmihalyi [27] illustrated in Figure 4. This theory describes a mental state of full focus and immersion in one activity.

Therefore, is is also desirable to *vary* the difficulty inside the flow region, where people are neither underchallenged nor overchallenged. Even failures are desired, because it improves the experience of mastering the challenge thereafter. Of course, it should also vary the challenge itself to prevent doing the same thing all over.
There is one more thing games are really good at: give *juicy feedback*, something often lacking from real-life. The term "juicy" means that the feedback is "fresh" and encouraging as well as, in contrast to real-life, a gamer has the possibility to get feedback at any given state of the game.
Although, there is a danger in the part of competence by

[7] http://stackoverflow.com
[8] progresswars.com

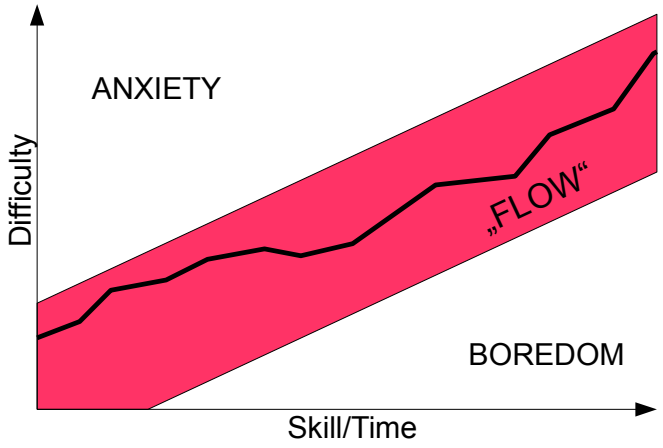


Figure 4. The Flow theory: As long as the difficulty of the challenge is within the flow, poeple are neither overchallenged nor underchallenged [21].

emerging *unintended behaviors* trying to avoid the challenge. One example is "Mayor Maker" for Foursquare, a tool which automatically checks in and out of places the user is just driving by.

Therefore, gamified applications have to address the user's competence need by offer tasks in a challenging and interesting way, while handing "juicy" feedback.

Recap this paragraph:

- Provide interesting challenges
- Provide clear, visual, varying, and well structured goals
- Provide jucy feedback
- Beware of unintended behaviors

### C. Autonomy

Most games are a *voluntary activity* and the choice to play is intrinsic. Following the "overjustification" effect [23], it can be dangerous if extrinsic rewards, like cash incentives, are used, especially in contexts of work. Therefore, if there is a "if..then" reward, people will tend to realize that they are *losing their autonomy* and being controlled, which is in general a really de-motivating experience. Furthermore, *devaluating the activity* is another point to beware of when using extrinsic rewards. For example, if there is a sweepstake from a service and the condition to take part is to re-tweet it, the signal is that the service is not good enough. For that reason, the people would not autonomously re-tweet it to their friends.
However, there are methods to deal with that. Like customer services shows that *shared goals, but individual pursuit* works quite well. As well as using informational feedback rather than controlling feedback and using unexpected rewards.

Wrapping it up:

- Play is voluntary
- Beware of losing autonomy


Figure 5. Choose personal goals for your own contract of this month. "DO's" are visualized in greed, while red is the color for "DON'Ts".


Figure 6. The chosen goals can further be customized to match personal interest as well as influence the level of difficulty.

- Beware of devaluating activity

### D. An example of a gamified application

Now let's have a look at healthmonth.com as one specific example. It is a pervasive health application which trys to help people live healthier.

*1) Relatedness:*

*a) Personal goals:* After the registration, the first task is to choose *personal* goals from a list of "DO's"(green) and "DON'Ts" (red). These goals can further be *customized* to range from easy to difficult and match the user's interest, see Figure 5 and Figure 6.

*b) Connect to a meaningful community of interest:* The gamer is connected to people who are also interested in enhance their health and the set of teammates is further determined on the difficulty of the chosen goals. Therefore, you always play with or against people on an almost equal level of interest. This placement can be seen in Figure 7.

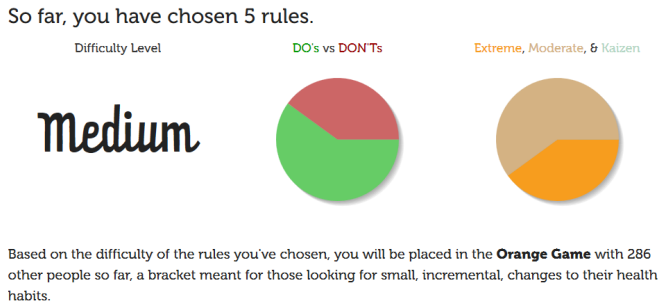


Figure 7. There is always a feedback about your current state as well as at this point you see where you would be placed depending on the currently chosen personal goals.


Figure 8. This shows a contract with the chosen goals. The gamer has to sign this contract to start the game.

Additionally, it is also possible to heal your teammates to help them if they missed a goal. This improves the importance of playing together, because it is not possible to heal oneself.

*c) The meaningful story and the social context meaning:* There is no fictional story, just the real-life and the shared monthly recap of the behaviors. While, healthiness is a desirable goal within the society, a player has to consider that his information can be watched by others. Therefore, it is possible to use the service anonymously.

*2) Competence:*

*a) Provide interesting challenges:* As the goals are determined by the users themselves, it is up to them to come up with interesting challenges. In Figure 8 you can see a contract, which is set up and has to be signed by the user to start the game. There are also free spaces to define own rewards or punishment. Nevertheless, the difficulty also depends on previous habits.

*b) Provide clear, visual, varying, and well structured goals:* As it can be seen in Figure 9 the goals are clear and visual as well as there are some additional incentives like team points and extra points. However, the variation as well as the structure is mostly in the hands of the user by defining the goals for each month. It could help to put some more incentives to increase the difficulty as well as points gained over time.

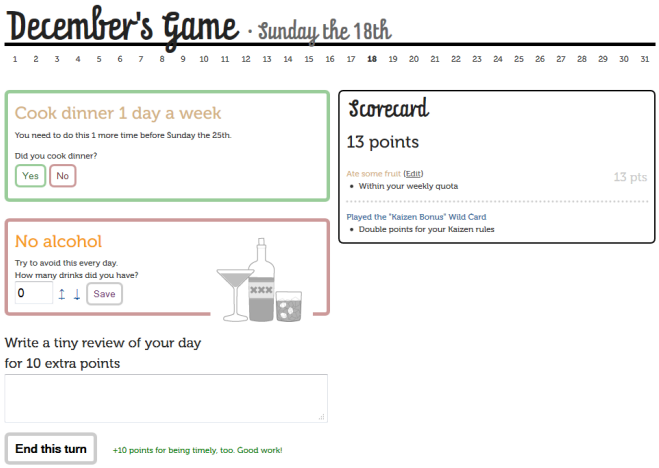


Figure 9. When you play healthmonth, you have to do daily turns, where you describe the progress of your goals.

*c) Provide jucy feedback:* You will get feedback all over the place, like it is visualized in Figure 7. Even though the feedback is discreet, it is enough and visualized well.

*d) Beware of unintended behaviors:* This is one of the most difficult parts to handle in the context of healthmonth.com, because there is no guaranty that a player is faithful in his tasks. Therefore, the whole concept is built on self-assurance as well as the trust within the community. Furthermore, there is no real benefit from cheating themselves.

*3) Autonomy:*

*a) Play is voluntary:* All of healthmonth.com is firstly intrinsic. Otherwise it would not have a real attraction.

*b) Beware of losing autonomy or devaluating activity:* However, there is a superior socio-economic principle which could be denoted as extrinsic. Therefore, there is the danger that people start to realize that. However, the result would be simply to stop playing the game.

## IV. The Future of Gamification

The chapter is a discussion about the future of gamification between hope, thread and hype.

One of the best-known proponents is Jane McGonigal. She is a game designer, researcher and author, specializing in pervasive computing and alternate reality games. One of the main aspects from her book "Reality Is Broken: Why Games Make us Better and How they Can Change the World" [7] is that for solving the world's problems it is important that people play more games and not less. One of the games which substantiate her argumentation is "World without Oil"[9]. Where, the gamers have to image that they are living on earth without oil and find out how they can survive. The players created podcasts on how they are living and helped each other on developing strategies for that. The second game is "Superstruct"[10] in which over 8000 people

[9] worldwithoutoil.org
[10] superstructgame.net

have tried to find superstructs to save mankind. The players have to develop ideas to hold superthreats like ravenous, power struggle or outlaw planet. However, there were no particular guidelines how such superstruct should be formed. Both games have the principle of "Play it - before you live it".

Of course, there are threats in the topic of gamification. One of this threats is the "Vision of the Gamepocalype" introduced by Jesse Schell [2]. In his vision there is a possible future, in which you play a game in every second of your life. For example: if you brush your teeth you will get +10 points, if you look at advertisement: +20 points, if you are eating the right cereals: +30 points and so on. This also goes along with Gabe Zichermann, he said: "games are the only force in the known universe that can get people to take actions against their self-interest, in a predictable way, without using force" [28]. Additionally, there is a huge privacy problem. Because all over, these gamified applications acquire a lot of personal data and in the way of achievements, presented in the internet, this data is also permanent.

## V. Conclusion

A huge discussion about gamification started in late 2010. For that reason, this overview about the current state of gamification is just able to present a small part of the whole discussion. Nevertheless, the definition introduced by Deterding et. al. is able to isolate the phenomenon of gamification from previously introduced areas of research as well as set it in relation to it [10]. While, their definition seems fitting, it will be probably just the start and it has to be further enhanced.

Since, the definition just says, that there are game design elements which should be used in non-game contexts. There is still much room how these game design elements should behave and look like. Therefore, the second part mainly discussed important game design elements and associated them to the "self determination theory", used as the base of a good gamified application. Whereas, there are maybe some applications which have success outside this spectrum, most of the current successful gamified applications can be substituted to fulfill these three points: competence, relatedness and autonomy.

McGonigal is one of the first researchers who examined the question what games can offer for the society in particular. Although, serious as well as pervasive games have a long tradition in game design, this has opened a new perspective. The crowdsourcing potential is currently used in a lot of gamified applications like DNA matching games foldit or phylo.

Finally, following the point about losing autonomy made before, the exaggerated vision from Schell and Zichermann is unlikely to happen. However, there are of course threats of privacy and "over gamification". Nonetheless, Schell makes a point that the hype about gamification is based on the fact that society itself changed into a more pleasure based society. Thus, gamification has its rights in the current discussions. However, for a good academic summary the hype has to cool down before and proper scientific studies about the benefits as well as the side-effects of gamification are needed.

## References


[1] F. Khatib, F. DiMaio, F. C. Group, F. V. C. Group, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popovic, M. Jaskolski, and D. Baker, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *nature structural & molecular biology, advance online publication*, 2011.

[2] J. Schell. Visions of the gamepocalypse. [Online]. Available: http://bit.ly/bT62k3

[3] D. Helgason, *2010 Trends*, Blog, Unity Technology Blogs Std. [Online]. Available: http://goo.gl/AZ4vm

[4] T. Chatfield, *Fun Inc.: Why Gaming Will Dominate the Twenty-First Century*. Pegasus Books, 2011. [Online]. Available: http://books.google.de/books?id=aYq7Fp79kzYC

[5] G. Zichermann and C. Cunningham, *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*, ser. O'Reilly Series. O'Reilly Media, 2011. [Online]. Available: http://books.google.de/books?id=Hw9X1miVMMwC

[6] D. Flatla, C. Gutwin, L. Nacke, S. Bateman, and R. Mandryk, "Calibration games: Making calibration tasks enjoyable by adding motivating game elements," in *UIST '11: Proceedings of the 24th annual ACM symposium on User interface software and technology*, Santa Barbara, California, USA, 2011, pp. 403–412.

[7] J. McGonigal, *Reality is broken: why games make us better and how they can change the world*. Penguin Press, 2011. [Online]. Available: http://books.google.de/books?id=yiOtN_kDJZgC

[8] B. Reeves and J. Read, *Total engagement: using games and virtual worlds to change the way people work and businesses compete*, ser. Harvard Business School Press. Harvard Business Press, 2009. [Online]. Available: http://books.google.de/books?id=hbFSia7bdUoC

[9] S. Deterding, K. O'Hara, M. Sicart, D. Dixon, and L. Nacke, "Using game design elements in non-gaming contexts," *CHI 2011*, 2011.

[10] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "Gamification: Toward a definition," *CHI 2011 gamification workshop*, 2011.

[11] ——, "From game design elements to gamefulness: Defining ªgamification," *Mindtrek 2011*, 2011.

[12] R. Caillois and M. Barash, *Man, play, and games*. University of Illinois Press, 2001. [Online]. Available: http://books.google.de/books?id=bDjOPsjzfC4C

[13] K. Alfrink, *New Games For New Cities*, Presentation, FutureEverything Std., May 2011.

[14] C. Abt, *Serious Games*. University Press of America, 1987. [Online]. Available: http://books.google.de/books?id=axUs9HA-hF8C

[15] I. Bogost, "Persuasive games: Exploitationware," *Gamasutra*, p. 4, May 2011. [Online]. Available: http://goo.gl/88969

[16] B. Sawyer and P. Smith, *Serious Games Taxonomy*, Presentation, GDC 2008 Std., February. [Online]. Available: http://goo.gl/xc1ZX

[17] J. Hamari and V. Eranti, "Framework for designing and evaluating game achievements," in *Think Design Play: The fifth international conference of the Digital Research Association (DIGRA)*, C. Marinka, K. Helen, and W. Annika, Eds. Hilversum, the Netherlands: DiGRA/Utrecht School of the Arts, September 2011, p. 20.

[18] M. Montola, T. Nummenmaa, A. Lucero, M. Boberg, and H. Korhonen, "Applying game achievement systems to enhance user experience in a photo sharing service," in *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, ser. MindTrek '09. New York, NY, USA: ACM, 2009, pp. 94–97.

[19] K. Huotari and J. Hamari, "ìgamificationî from the perspective of service marketing," *CHI 2011*, 2011.

[20] J. Schell, *The Pleasure Revolution: Why Games Will Lead the Way*, Presentation, GoogleTechTalks Std., November 2011. [Online]. Available: http://youtu.be/4PkUgCiHuH8

[21] S. Deterding, *Meaningful Play*, Presentation, Google Tech Talk Std., January 2011. [Online]. Available: http://youtu.be/7ZGCPap7GkY

[22] A. Kohn, *Punished by rewards: the trouble with gold stars, incentive plans, A's, praise, and other bribes*. Houghton Mifflin Co., 1999. [Online]. Available: http://books.google.de/books?id=-fFsZWg-JfAC

[23] M. Lepper, D. Greene, and R. Nisbett, “Undermining children’s intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis.” *Journal of Personality and Social Psychology*, vol. 28, no. 1, p. 129, 1973.

[24] E. Deci and R. Ryan, *Intrinsic motivation and self-determination in human behavior*, ser. Perspectives in social psychology. Plenum, 1985. [Online]. Available: http://books.google.de/books?id=p96Wmn-ER4QC

[25] S. Rigby and R. Ryan, *Glued to games: how video games draw us in and hold us spellbound*, ser. New directions in media. Praeger, 2011. [Online]. Available: http://books.google.de/books?id=H3MipRj4Id0C

[26] R. Koster, *A theory of fun for game design*, ser. Paraglyph Series. Paraglyph Press, 2005. [Online]. Available: http://books.google.de/books?id=GQpQAAAAMAAJ

[27] M. CsÌkszentmih·lyi, *Flow: The Psychology of Optimal Experience*, ser. P. S. Series. HarperCollins, 2008. [Online]. Available: http://books.google.de/books?id=epmhVuaaoK0C

[28] G. Zichermann, *Fun is the Future: Mastering Gamification*, Presentation, Google Std., October 2010.

# Body Area Networks

Jonathan Sondershaus
Ulm University
Ulm, Germany
Email: jonathan.sondershaus@uni-ulm.de

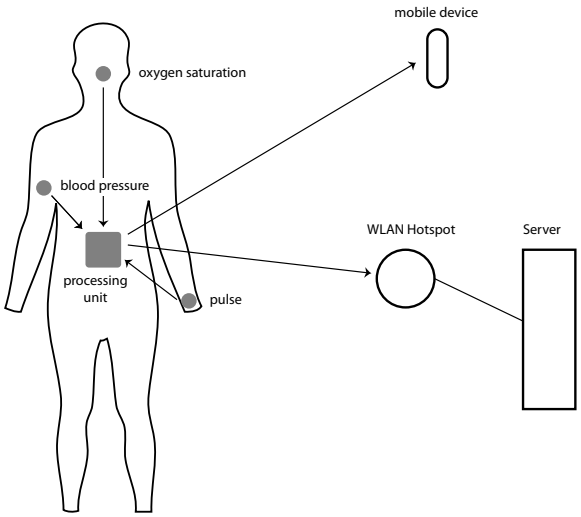


Figure 1. Architecture of a Body Area Network

***Abstract*—Body Area Networks can improve every day life by connecting different devices together, collecting and converting data. But they have also address several important issues like privacy and security of user. Energy efficiency is also very important and can be influenced by memory, batteries and connection protocolls.**

## I. Introduction

Everyday life is supported by many different devices such as MP3-Player, celluar phones, clocks measuring pulse and so on. *Body Area Networks* are able to connect all devices, to get a system of support. Figure 1 shows a simple *Body Area Network* which can be used to collect health information to observe a patient. Sensors can be implanted inside the human body or can be weared. [1]

### A. Scope of Application

In the following four possible application areas of *Body Area Networks* are described.

*1) Fitness:* A person wants to train his/her body. So he or she goes to a fitness center. To improve the personal training program sensors for blood pressure, pulse, oxygen saturation, temperature are placed on and inside the body as shown in Figure 1.

Beside the sensors a data processing unit is installed on the body to collect data from sensors, temporarily store them and to preprocess them.

Another two devices are installed on the body. One able to connect to a wireless lAN or different network to send data to a server where the information is processed and the training program will be refined. And the other one to connect to a mobile device to deliver realtime data to the user of his current vital signs so the person can self influence the current training program.

At all those parts technology creates problems, risks and issues that have to be discussed.

*2) Health care:* Body Area Networks cannot only support a fitness program, but also support other activities in everyday life. Healthcare is also a possible application area for BAN and WBAN.

For example: A patient is hospitalized with suspicion of a heart attack. During examination in the hospital the suspicion is substantiated. After clinical therapies the patient is send home, but needs to be monitored. Usually that means a change in everday life for the patient. All sensors and devices have to be carried in pockets or backpacks. In order to not influence his everyday life too much a Wireless Body Area is installed. Sensors for blood pressure, arterial flow, oxygen saturation are getting implanted inside the body and sensors to monitor physical trends like falling and stumble are placed on the body. Some actuators to supply the patient with medicaments are also installed. So the patient can live the everyday life without taking too much attention on how to transport all things. [1]

Many health therapies need long term monitoring of patients. But this monitoring costs a lot of money, because patients typically have to stay in hospitals for a long time. Aside from the time aspect, there is no possibility to build an everyday life atmosphere around the patient. So a lot of sensors are placed on the body and processing units are carried inside a pocket, bag-pack or something else monitoring the patient outside hospital. But all these things influence everyday life. [1]

There are many other cases where wireless body area networks can help to improve monitoring. Patients do not need to go to ambulant examinations as often, every health status can be easily accessed by patient and doctor through remote devices.

With Body Area Networks it is possible to connect the devices without distracting the patient and even get better and faster information about the current health status because information can be processed in realtime and transmitted to the hospital through internet.

*3) Military use:* Even a soldier can get advantages from Body Area Networks. Physical status of an soldier is essential for his commander. Collected information about the body and

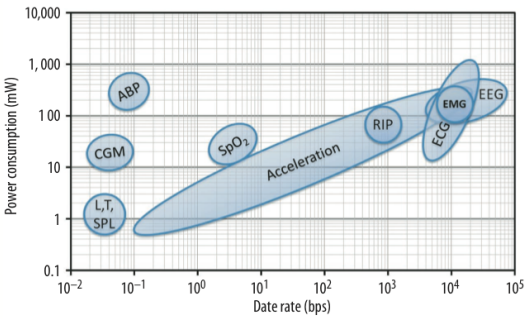


Figure 2. from Body Area Sensor Networks [1] Data rate compared to power consumptions in healthcare sensors

the environment can be used to prevent soldiers from possible risks.

*4) Everyday life:* There are even more devices that are possible to connect together. Having to deal with tangled headphone cables can be avoided by wireless body area networks. A lot of mobile devices, able to play songs, have the abilities to connect to devices through Bluetooth or Infrared. Adding some sensors to the body to measure pulse and blood pressure could even provide the ability to choose from different playlists to present songs that fit to the current mood.

## II. Components

### A. Sensing

First step of collecting information is sensing. Small devices on the surface of the body or inside are collecting different information such as blood pressure, oxygen saturation and others.

These sensors must not harm the person or influence everyday life. The sensors have to be small and energy costs have to be reduced to guarantee a long lifecycle. The location where the sensors are placed can also be a big problem. Misplaced sensors can collect wrong or divergent data.

### B. Actuators

Actuators are small medical devices which can be a medicine supply for the body. Combined with sensors and a processing unit these devices can supply the body with medicine to prevent bad health status. For example if a processing unit gets information from a sensor measuring blood flow inside coronary blood vessel and sending the processing unit that there is a possible risk of a heart attack, the processing unit can send an actuator a signal to supply the body with medicine to thin out the blood.

### C. Processing Unit

Because of the small size of sensors they are not able to store the collected data or process them. So they send all collected information to a data processing unit which is also placed on the body to prevent long data transmissions. The fundamental task of the processing unit is to preprocess collected information, filtering unnecessary information. It is even possible to store them temporarily in case of a missing connection to a network nearby.

The main problem in collecting and preprocessing the information is how data is collected. On the one hand, there is continuous information like body temperature and oxygen saturation, but on the other hand there is also discreet information like blood pressure.

All collected information has to be stored temporarily and has to be synchronized by the processing unit to get the correct information. Information from multiple sources can only be interpreted when it can be compared to stored information like blood pressure. One value cannot represent the health status of a patient, only compared to references.

Much data is not measured continuously. Some sensors only send data, when an event happened or when a threshold has passed. Especially when a lot of sensors are placed it is important to synchronize all information. If the sensor for pulse frequency is placed on the same arm as the sensor for blood pressure the second sensor can influence the measurements of the first.

To prevent sensors to cost to much energy the period of data mining should be reduced. Not every information has to be collected every second.

### D. Data storage

Data storage devices are getting better and smaller. For a body area network it is necessary that the storage does not consume to much energy. But because of the energy saving issue and the fact, that devices should be as small as possible the data storage can not be so big. Therefor the data mining process and preprocessing of data has to be very good, to prevent of saving unnecessary information.

### E. Communication

After collecting all information there are two destinations where the data has to transmitted to. On the one hand the person wants to get a feedback of what has been collected. In the example of fitness center the person wants to know the current pulse and blood pressure. On the other hand all information has to be transmitted to a server to improve the training program.

There a lot of possible communication types that can be used. Most people have a mobile phone, which is able to connect to Bluetooth, infrared or wireless LAN devices. So they can be used to offer a realtime feedback for the person. These connection types make it also possible to connect to a WLAN to transmit the collected data to a local server of the fitness center or send them through internet to a server.

### F. Power supply

All devices needs electricity to work. Devices carried outside the body can use replaceable batteries and they can be recharged separately. But sensors inside the body need batteries too. It is not possible to change batteries of those sensors after few days without extracting them from the body. So either the batteries installed guarantee a long time power supply or they have to be recharged from outside the body.

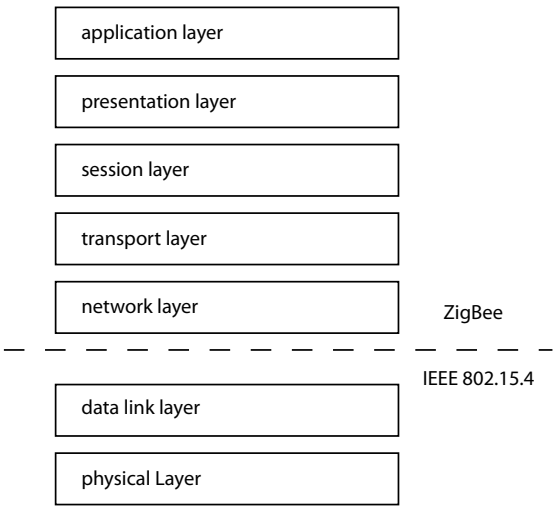


Figure 3. OSI-Model

There are medical products which are implanted inside a human body and can be recharged from outside like a cardiac pacemaker. Those devices are also systems which have some sensors, processing units. actuators and a power supply. Those batteries supply the pacemaker for several years.

## III. Connecting everything

The connection between the processing unit and the environment differs from the connection of processing unit and its sensors. In the connection between processing unit and a server the processing unit has to adapt to local conditions.

### A. Connecting sensors and processing units

Main task of connecting sensors and processing units is a secure connection which does not cost a lot of energy and can be easily established. Because of the small distance between devices a low power and low intensive signal can be used. The IEEE 802.15.4 Standard defines such a standard for Wireless Personal Area Networks. It contains the two lowest layers of the OSI Standard, as shown in Figure 3.

This protocol is also the base concept of another Standard, developed by an association of more than 230 companies and organizations. It is defined for the use of a sensor network. Whereas the IEEE 802.15.4 standard is only defined on the lowest two layers, the ZigBee-Standard contains all seven layers (see Figure 3). IEEE 802.15.4 and ZigBee mostly use the free ISM-Band to connect devices.

Using a standard is important to develop new devices fitting in existing networks. Knowing the problem of mobile phone recharger, this problem should not produced inside body area networks. One advantage of using a technology based on the OSI-model is, that connecting the devices together works nearly the same than connecting to other devices around the body area network, although there have to be other security standards.

Another possible technology is the use of Bluetooth. The high energy costs of a Bluetooth connection although they have been improved recently [2].

Using the IEEE 802.15.4 Standard or ZigBee it is possible to implement a secure connection between the devices. This is important to prevent attacks to the body area network.

One problem can possibly appear by using this standard. Mostly the ISM-Band is used to implement a connection with the IEEE 802.15.4 or ZigBee standard. This is an ultra-wideband connection with several possible channels. But when there are too many devices sending information at the same time it is hard to separate them.

If no wireless connection is necessary it is even possible to install sensors inside special textiles, which contains wires to connect the devices together.

### B. Alternate connection type

Using the human body itself to establish a connection between the devices is another possible way. IDENT [3] is a company specialized on using the human skin to transmit data. Together with Hochschule München they have invented the Skinplex-Technology [4]. With this technology it is possible to connect sensors and processing units without letting the information outside the human body. The human skin is possible to transmit electronic signals. This technology is also used by lie detectors. Many attributes like thickness, temperature and wetness influence the signal. Using this technology it is possible to transmit personal data without make it accessible by third parties.

There can be even more alternate connection types to be developed in the future.

### C. Outside Body Area Network

The easiest way to connect a Body Area Network with its environment is to use a reachable WLAN. During the years a lot of public hotspots inside public buildings have been

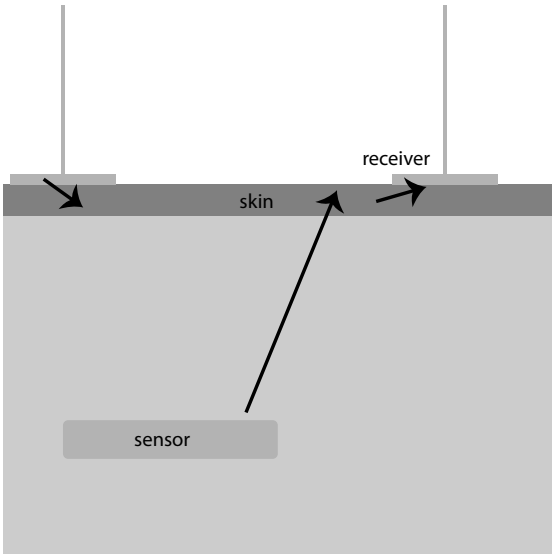


Figure 4. Transmitting data through human body

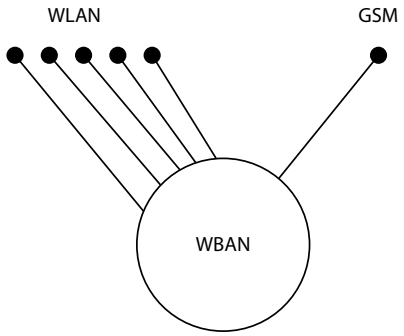


Figure 5. Connections to environment

established. So the processing unit can connect to those hot spots to transmit information through internet to a server. The common HTTP-Protocol can be used to transmit all necessary data. It is easy to establish a secure connection with HTTPS.

This connection type can also be used by several other devices that have to be plugged into the network. For example, if a body area network is used to measure health status, it can easily add other devices to improve the fitness program. Adding a devices counting the steps a person makes during the day and combining those informations how many steps have been made every minute. These additional devices can be integrated during workouts and can easily be removed after workout.

In fact the body area network should not connect automatically to all reachable WLAN hotspots. First of all not every body area networks needs to be connected to the internet. There are several data that can be stored inside the network and can be synchronized later.

Other possible connection types are GSM, Bluetooth and Infrared. Technology of choice is the one that is reachable in application area.

When Wireless Body Area are used to improve the personal training plan, they only have to connect to a WLAN inside the fitness center and perhaps at home to synchronize reference values for the sensors. In military use it is possible to use GSM, because outside buildings there normally exists no Wireless Local Area Network and the information that are collected often needs to be transmitted in realtime.

## IV. Synchronization of Information

Many information sources deliver their data in different time periods. But often the information has to be presented in a synchronized form. To get those information the processing unit has to collect all data from the sensors and combine it.

It is even possible that more than one processing unit is installed inside the body area network. The system must provide data sharing between them. The information stored

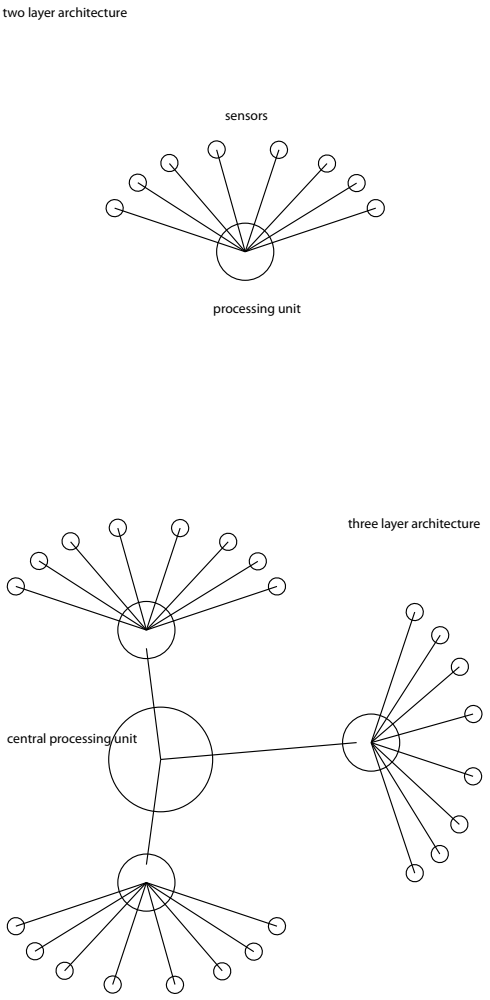


Figure 6. Possible architecture of two or three layer architecture

from all units can be independent or redundant. In the later case, stored information has to be compared.

For example there can be an architecture implemented with basically two layers (Figure 6), the sensors and a processing unit, but on the other hand it is possible to have three or more layers, processing data in different steps.

## V. Feedback to user

If a device or system does not deliver any feedback to the user working with it (Figure 7), the user does not know whether the system is doing the right things. So giving feedback is essential for the acceptance of body area networks.

In health care this feedback can be easily produced by delivering additional information to the patient wearing the body area network. Patients should be able to get realtime data of current health status.

A user of a body area network has to be informed about all information that are stored and produced inside the network. This does not mean that every information must be provided in realtime. These data also contains information of who else is able to access the personal data. In case of a medical

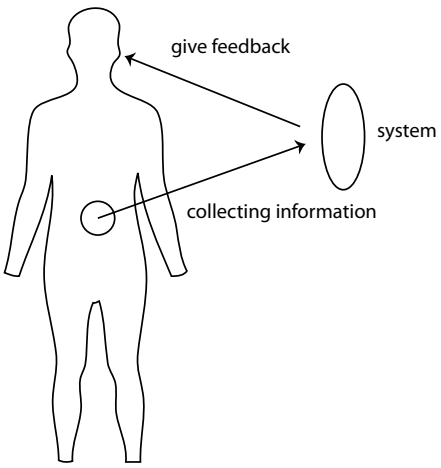


Figure 7. System collect information and gives feedback

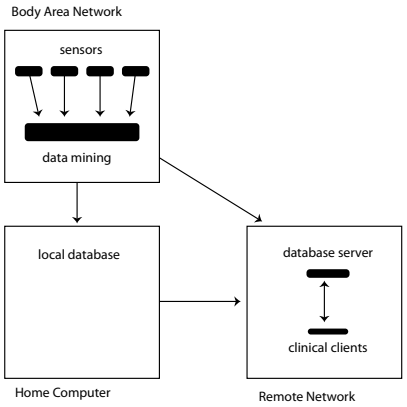


Figure 8. possible connection configuration

application these persons would be all staff that are integrated in the medical therapy.

Warnings of possible decreasing health status is an important feedback, the user have to be able to terminate this warning. Such warnings can be produced because of a malfunctioning sensor or other things which happen around the person. Not every increasing pulse and blood pressure is a sign for a critical situation. This is why the person must be able to influence the processing of data.

## VI. Seamless Configuration

Seamless Configuration is not easy for Wireless Body Area Networks, because of signals transmitted through air. They can easily fetched by third party persons. There are applications, like connecting earphones with MP3-Player, that do not need high security standards, but there are several applications where this is an essential task.

If devices establish connections to all access points around, the user will not able to control where his data will be transmitted. The devices should only connect to other devices that can be trusted. It is similar as the problem with a webserver through the internet.

This problem exists only in wireless body area networks. Systems that are based on sensors, connected through wires can be configured by default. There is no security problem of personal data. Only the time interval of collecting information has to be configured to fit personalized circumstances. But this can be done by using a remote device, either an application run on a computer or a mobile device which both are connected to the body area network.

Inside a wireless body area network there have to be a trust management. Because of other body area networks from other persons around there will be the possibility of connecting to the wrong device and producing false information send to the server. If there is no manual configuration it would be easy to scan wireless body areas by sitting next to the person, tracking all informations about health status and personal data stored on the devices.

A possible connection configuration is illustrated in Figure 8. Whereas the connection between the units inside a body area network without using wireless technology and the connection to a home computer is very similar by using wires, the seamless connection to remote networks is not possible because of the risks of an attack of third parties.

## VII. Safety

Sensors can deliver wrong data. It is not easy to detect whether the delivered information is correct or wrong. But how can wrong data deliver can be avoid. To understand that it is necessary to have a look at possible causes of false informations.

First of all sensors can be misplaced. But that should not happen during implantation inside the body. Sensors that can be worn outside the body can easily be put back in place by the person who wear them.

A bigger problem are the risks that all electronic devices have. Because of the use of electricity, sensors have to be prevented from short-circuiting and sending electricity inside the human body and produce injuries.

Energy saving is very important for every device in body area networks. Only long battery durations make it possible to use the networks independent from several maintain sessions.

Actuators that supply medicine are under the same constraints as other medical products are. This means also the software that controls these devices. A malfunction inside can badly harm the person using the body area network.

Every body area network which monitors critical tasks needs a backup function. For every function there has to be a possible alternative solution. [5]

## VIII. Privacy

All collected data is very personal. They must not be accessible by a third party. Therefore they have to be protected.

Whereas the connection between the sensors and the processing unit is relatively risk-free because of the short distance between them, the connection between the processing unit and a mobile device or server has to be as secure as possible. Used protocols between them must have a possibility to encrypt the data and transmit them securely. Authentication-protocols can help to establish a secure connection between the Body Area Network and the server. [5]

Although a good balance between privacy and longtime use of devices should be the goal, it is more important to guarantee a good privacy handling than saving energy. Energy can bis recharged, harming privacy can not.

In case of using the Skinplex-Technology [3] there is no need for encryption the data between sensors and processing unit. But data has to be encrypted for transmission outside the Body Area Network. All other technologies have to connect to each other through a secure connection to prevent attacks from third party.

Authentication is important for every connection that has to be established, between devices or between the WBAN and the network around. A person sitting next to another person in train should not be able to collect data of the mobile device storing business information. This information must be protected. Figure 9 illustrates the possible overlapping of different body area networks. If the connection between the devices is wireless an authentication is necessary to prevent wrong transmitted data.

Establishing connections to the environment goes along with risking to be trackable. When a third party can get the information where the person wearing the Body Area Network has been connected to, this person can be tracked. Although this must not be a realtime information, but it can be used to collect longtime data and get information about the persons everyday life.

If mobile devices are connected to the wireless body area network and there is no trust management it is easy for a person sitting nearby to collect all information stored on several devices.

## IX. Conclusion

Body Area Networks can be a good support in everyday life. In health care, they can help people live a normal life, but be able to react to critical health status as soon as possible. Those networks can help making life easier but should not be the only source for information and support. Placing sensors which measure several vital functions can be used to get a better picture of a human health status and to research the cause of deceases. It is also possible to get better data of how good a medical therapy or a fitness program works or if those therapies harm other vital functions.

Although technology has been developed so far, there are possible improvements necessary to provide a secure and comfortable body area network that can be used easily like the connection between devices and computers through USB.

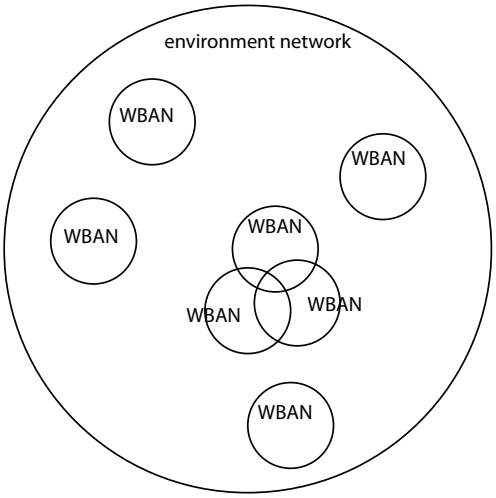


Figure 9. Overlapping networks

## References

[1] M. Hanson, H. Powell, A. Barth, K. Ringgenberg, B. Calhoun, J. Aylor, and J. Lach, "Body area sensor networks: Challenges and opportunities," *Computer*, vol. 42, no. 1, pp. 58 –65, jan. 2009.

[2] C. Otto, A. Milenkovic, C. Sanders, and E. Jovanov, "System architecture of a wireless body area sensor network for ubiquitous health monitoring," *Journal of Mobile Multimedia*, vol. 1, pp. 1–20, 2006.

[3] Skinplex. [Online]. Available: http://www.ident-technology.com/

[4] P. D. C. Rapp, "Datenübertragung über die haut - skinplex technologie," 2008. [Online]. Available: http://www.forschungskoop.de/60_pdf/pi-beispiele/pro_inno_227.pdf

[5] S. Warren, J. Lebak, J. Yao, J. Creekmore, A. Milenkovic, and E. Jovanov, "Interoperability and security in wireless body area network infrastructures," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, 2005, pp. 3837 –3840.

# Flying Ad-Hoc Networks

Michael Müller
Ulm University
Ulm, Germany
michael-4.mueller@uni-ulm.de

***Abstract*—The research presented here aims on releasing a swarm of Micro-Air-Vehicles from a base station. The swarm organizes itself to search for a user in a large area. Once the user is found a communication link must be established and maintained from the base station to the user.**

**The two concepts presented here aim on building a swarm that only relies on local communication with neighbors. No global positioning is available. Analogous to real swarms the sensors only provide data about the immediate environment of an individual.**

## I. Introduction

Swarming Micro Air Vehicles (SMAVs) can be used to provide a large area with a communication network. The research presented in this article aims on releasing a swarm of autonomous MAVs who organize themselves to establish a wireless communication network. We specifically address the issue of establishing a link from a base station to a user startion. The MAVs hereby are explicitly not equipped with any Global Positioning Systems (GPS), cameras, laser range finders or radars. Sensors which are dependent on the environment have the disadvantage that they are expensive in terms of energy, size, cost and weight. Additionally GPS is unreliable or not necessarily possible in disastrous areas. For positioning the GPS receiver must read four satellites simultaneously. To get the most exact localization, a line-of-sight with the necessary satellites is ideal. In confined spaces like a city with tall towers or a closed hall this is not necessarily possible. Since it can't be ensured that GPS always is exact, this effectively limits the deployability of the system [1].

Instead the researchers aim for minimal platforms which are cheap, safe, light-weight and easily deployable. The MAVs hereby are equipped with sensors that enable them to sense their immediate environment. Those sensors provide heading, altitude, speed and possibly communication with their direct neighbors.

Examples for scenarios where a communication link spanned by a flying swarm is needed include disaster areas where the topology of the compound does not allow to install base stations in a short amount of time. Other examples are disasters where floods prevent a stable power supply.

The typical scenario addressed in the research presented in the following sections, is a group of rescuers who arrive at a disaster scenario and rapidly have to set up a base station and a communication network. MAVs are then launched from the base, one after the other, and must organize themselves to find rescuers or victims. Once found a communication link has to be established and maintained until no longer needed.

Flying robots are especially fitted for such scenarios because of their easy deployability and their ability to spread out over difficult terrains like flooded areas or debris.

A connection between rescuer and base station is helpful for exchanging rescue-related informations, positions of victims or photography.

The fundamental role that modern information and communication technologies (ICT) play in rescue operations, such as environmental and humanitarian relief operations, has been recognized in the last few years. In the following article we address the field of preservation of nature and human lives through utilization of modern technologies. Modern technologies can be used by search-and-rescue teams to gather a more precise knowledge base, resulting in an increased efficiency of operations [2].

## II. Swarm behavior

The research presented in this article aims on establishing a communication link from a base station to a user station, which is placed somewhere in the search area. To solve this task many Micro-Air-Vehicles (MAVs) are released into the air. Each of the MAVs has a wireless network link integrated. This way MAVs can communicate with each other and send packets. Through using Ad-Hoc routing protocols and each MAV as a node, a communication network can be established. This way a base station is able to communicate with the user station through the network.

Together the MAVs form a flying swarm with each individual acting on his own. Since each MAV only senses its immediate environment and has no knowledge of any global state, a swarm behavior has to be established. Meaning altogether the swarm works towards one common goal, though the individuals act on their own. This is called emergent behavior. Craig. W. Reynolds proposed several rules for enabling bird-like flocking within a swarm in 1986 [3]. The rules were derived from real bird-flocks and can be briefly summarized as: move with the same speed and direction as neighbors, avoid colliding with them and stay close.

Before a communication link to a user station can be established the swarm first has to search the area for the user station. Since the MAVs can only communicate with their direct neighbors the swarm forms a tight chain. Most of the experiments assume a rough knowledge of the direction of the user station. After the user station is found a communication link has to be established and maintained. Ideally the MAVs should arrange themselves in the best way to maintain the network. This way the packet delivery rate of the network gets

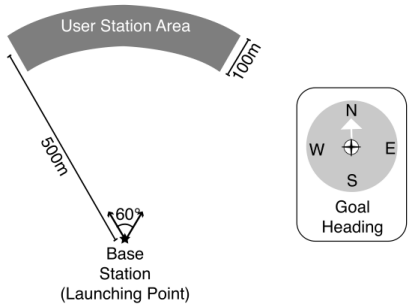


Figure 1. In this scenario environment a swarm should be able to find user stations placed within the User Station Area. The image is taken from [4].

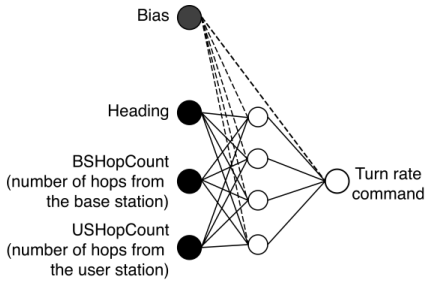


Figure 2. This neural network is used as a controller for the MAVs. The best fitting weights of the neural network are gathered using artificial evolution and extensive testing. The image is taken from [4].

optimized the longer the communication link is established. We will now present two basic approaches for finding and connecting to a user station.

The first approach uses genetic programming as a mean to generate controllers for the MAVs. Later the generated program is reverse-engineered and adapted to the specific situation.

The second approach we present, aims on modeling the swarm behavior on patterns observed and studied in nature. These bio-inspired techniques are used to build a model for the controllers.

## III. Artificial evolution

There is currently no proven methodology to deterministically design software with emergent behavior for a swarm which is directed towards a common goal. It is hard to use traditional programming techniques therefore, because problems of interest are not linear, stochastic or deeply enough understood. As an alternative approach the researchers within the SMAVNET project [4] [5] use artificial evolution to generate neural controllers for fixed-wing MAVs. An advantage of evolutionary approaches is the discovery of creative swarm strategies, which otherwise would not have necessarily been thought of.

### A. The MAV model

The researchers use a 2D simulation in which MAVs are launched every 15 $\pm$ 7.5 seconds within 50 m surrounding the base station. In a real life scenario the MAVs would get launched one after another.

The user station is placed in the area 500 $\pm$ 50 m away, in an angle of $\pm$ 30° from the base station. This means an approximate knowledge of the user station location is assumed to be known (see Fig. 1).

The MAVs in this model constantly fly at a speed of 14 m/s. One assumption of the communication model is that two agents are able to perfectly communicate if they are less than 90 m apart. Communication gets noisy from 90 m to 100 m. The probability of dropping a message increases linearly. If they are 100 m or more apart, communication is assumed to be entirely lost.

The agents can send two types of messages. Control messages have the purpose of coordinating the swarm and are broadcasted by each node every 50 ms. Data messages are used for the application of the communication (e.g. video, voice or data transportation). They are sent between the base station to the user station.

### B. The neural controller

For the artificial evolution a weighted artificial neural network (see Fig. 2) is used. Artificial neural networks can be used to model complex relationships between an input and an output. They can be described as a mathematical function $f : X \rightarrow Y$ or a distribution over $X$ and $Y$. The neural network for the MAV controller consists of several inputs:

*1) Heading:* A magnetic compass is used to determine the heading of the MAV.

*2) BSHopCount:* The minimum number of network hops from the MAV to the base station.

*3) USHopCount:* The minimum number of network hops from the user station to the base station.

The inputs were scaled to fit the range $[-1, +1]$ and as a neural function a hyperbolic tanget (tanh) was used. Synaptic weights were chosen in the range $[-4, +4]$ and coded on 8 bits. Additionally there are 4 hidden neurons in the neural networks. This was found to yield neural controllers with the highest fitness for the swarm controllers. It is noticeable that all inputs are exclusively derived from the immediate environment of the MAV (situated communication, [6]).

The neural network outputs a turn rate, which is adopted by the MAV. The speed of the MAV is not taken into consideration for the neural network, since it is assumed to stay constant.

15 independent evolutions were executed using a genetic algorithm [7]. For the first generation the genomes were initialized randomly. Selection was used to favor inter-agent cooperation. From the generated neural controllers a swarm of 100 individuals was formed. Each individual in the swarm was equipped with the same neural controller (a homogenous swarm). In real-life applications this would be an advantage since it enables interchangeable agents and allows for scalable systems. The fitness of each swarm was then evaluated for 10 user stations randomly placed within the user station area.

The top 10 performing controllers were then copied to a new population (elitism). The remaining population for this new swarm was generated by randomly pairing individuals

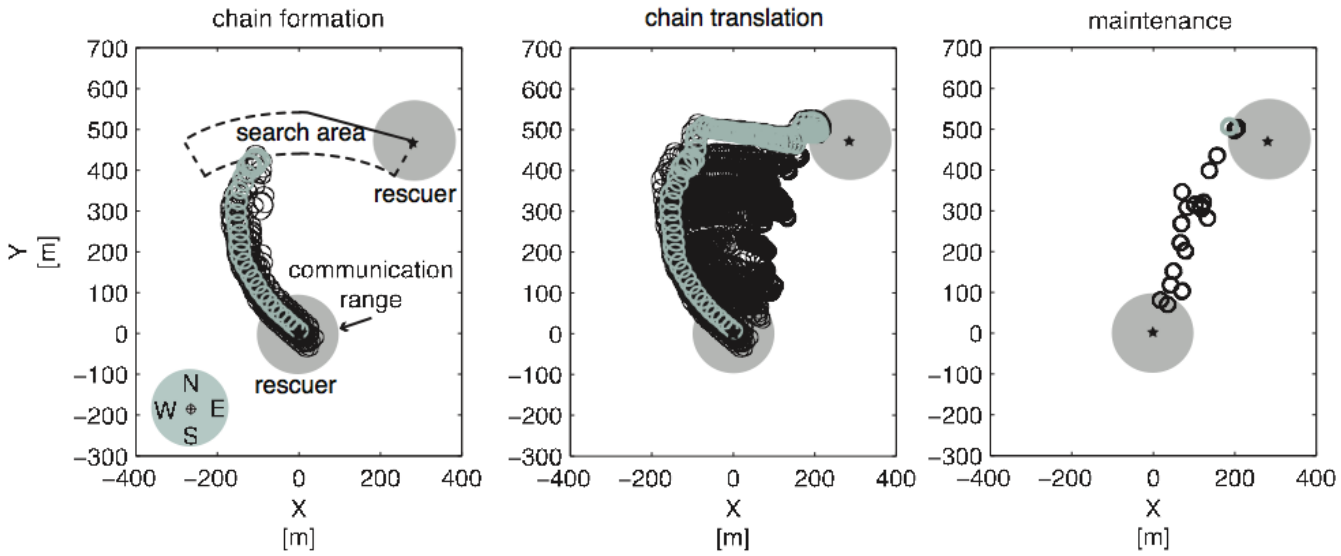


Figure 3. The first two images show the sweeping behavior of the swarm. The image on the right shows how the MAVs maintain a communication link to the base station. The light gray line shows the trajectory of the first launched MAV. The image is taken from [5].

within the first 30 ranks. The researchers favored controllers which were able to rapidly establish a connection and maintain it through the entire duration of the trial (30 minutes).

To achieve selection which favors the maintenance of a connection, the fitness $F$ of each swarm is measured by the mean connectivity between base and user station throughout the duration $T$ of the trial.

$$F = \frac{1}{T}\sum_{t=0}^{T} c(t)$$

Furthermore the connectivity of the network was regarded by introducing $c(t)$, the number of disjoint paths between base and user station. $c(t)$ also expresses the minimum number of MAVs that need to be removed for the communication link to break. As an additional action agents who lost contact to the base station for more then 30 s were assigned a fitness of 0.

After this phase, the best controller was tested with randomly placed user stations in the search area for 1000 trials. The connectivity of the best evolved controller can be seen in Fig. 4. The connectivity during the first couple minutes is 0 because the swarm has not spread far enough yet. But once a connection is established it is maintained throughout the remaining duration of the trial.

### *C. Analysis of the swarm behavior*

It is important to understand the evolved controllers. Only through a deep understanding it is possible to modify the controllers and rapidly adapt them to new scenarios. This way controllers need not to be re-evolved for each scenario. These principles can possibly be used in further research to hand design swarm strategies for MAVs. For this purpose the swarm behavior was analyzed to find the underlying principles responsible for the emergent behavior.

Several simple patterns were found to be responsible for the performance of the swarm. The MAVs do not fly straight. They rather fly in circular trajectories and periodically modify their turn rate. Though they fly in circles, this can be characterized as a flight in a global direction, since they adapt the turn rate. The trajectory results in shapes like Fig. 3. The maximum turn rate is limited to 100°/s. Agents are unable to make turns sharper than 18 m in radius. By modifying their turn rate the MAVs influence their direction and the speed in which they move towards one direction. Though our MAV model assumes a constant speed of 14 m/s the circular trajectories result in a small overall speed. This way MAVs who are already deployed and fly are able to wait for newly launched MAVs. MAVs who are further away from the base station adopt a negative heading angle. This results in a slightly curved swarm chain as seen in Fig. 3.

The swarm forms a tight chain that coherently moves from one side of the search area to the other. This results in the whole swarm scanning the area. Once a user station is found and a connection is established, the swarm reorganizes to maintain the connection. For this purpose the MAVs basically adopt a hovering behavior. Since the MAVs constantly move, they adopt the smallest, for their fixed-wing architecture, possible trajectory.

It was found that after MAVs disconnect from the base they adjust their turn rate to get back to a connected state,

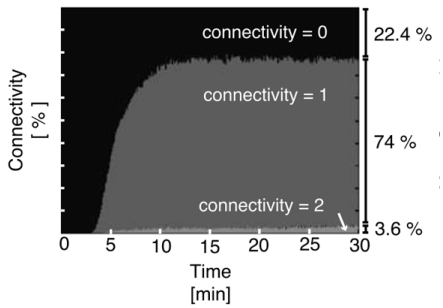


Figure 4. The connectivity of the best evolved controller over 1000 trials. Image taken from [4].

by tracking back. This behavior is further used to synchronize the swarm. After all MAVs are launched the whole swarm will eventually disconnect from the base station. At this point all MAVs perform a backtracking operation. This results in the whole swarm synchronizing on the heading. This way all MAVs display the same heading at at given time. This results in the swarm being able to move coherently from West to East. The connection and disconnection phases are used to align all MAVs.

This issue of synchronizing a swarm via broadcasted signals has also been addressed in one article on this seminar last year [8].

### D. Results

Through simulations the performance has been extensively tested and as a result the swarm was able to find more than 97% of the user stations placed within the search area. The swarm was able to maintain the communication network active throughout the remaining trial. All by just gathering informations via immediate sensors or through direct communication with neighbors. No global positioning was used. The researchers aim on being able to adapt the controllers to a variety of different scenarios in the future. The key concepts found within the artificially generated neural controllers can possibly be used to manually create controllers for a specific task.

## IV. Army Ants

In this section we present another approach for the same problem discussed before: finding, establishing and maintaining a communication link to a user station placed within a search area.

The approach presented here aims on using techniques observed in nature to build bio-inspired swarm controllers for MAVs. Extensive research [9] [10] found that different army ants species have evolved different raid patterns for different food distributions. The results indicate that the observed patterns maximize the amount of food brought back to the nest. The ants use this structure as an optimized mechanism for exploring and exploiting food resources [11] .

### A. The Ant Model

In the model discovered ants leave the nest at a constant rate. They navigate through a grid of Y-branches while constantly emitting pheromone. On each branch the ant chooses to turn either left or right. This decision depends on the amount of pheromone laying on the branch. The pheromone is also used as an indicator for the amount of ants who already chose to walk on a certain branch. This way the pheromone can be used to limit the number of ants deciding for a certain branch. The initial choices are made randomly, which explains the big difference in movement speed. The ants on the trails move very rapidly, they choose which branch to take very fast. Ants at the front on the other hand move much slower, with more hesitation. An explanation for this is that the explored trail is very well marked whereas the front is unmarked.

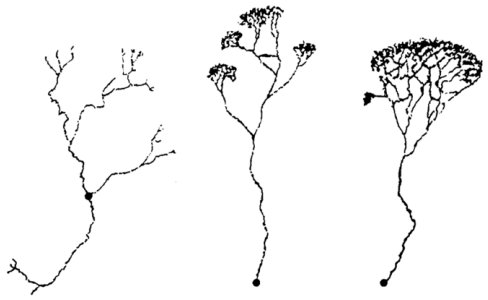

Figure 5. Foraging patterns of different army ant species, adapted to different distibutions of food sources. Each covering some 50 m x 20 m. The image is taken from [9].

Once an ant finds a point which contains food it takes an item and returns to the nest. On the way back it lays a greater amount of pheromone on the path. This way a yielding path is signalled. Eventually the pheromone laid on a path evaporates over time.

Fig. 5 shows different army ant species, each species hunts for different food resources. For example the *Eciton hamatum*, shown on the left, hunts for widespread, scarce food resources whereas the *Eciton burchelli* shown on the right largely feeds on a single, very large food resource.

### B. Searching the area

The idea of using pheromone trails for coordinating the actions of many MAVs in a swarm is described in [12]. The authors have designed a behavior for the MAVs that enables them so search the area. The problem of depositing pheromone trails could be solved by either depositing physical objects or chemicals. But depositing substrate in the air is very unstable due to the rapidly modifying environment. Depositing the pheromone on a virtual map is also not possible, since the MAVs only have local position awareness. The researchers used another approach instead: Virtual pheromone is deposited by separating the MAVs into two types: node-MAVs and ant-MAVs.

Node-MAVs form the environment in which virtual pheromone can be deposited. This is realized by the node-MAVs indicating the deposited pheromone for the left and right branch.

Ant-MAVs are then able to navigate through a grid of node-MAVs. Once they reach a node-MAV they select either the left or right branch based on the pheromone information which the ant-MAV contains. Once they don't find any node-MAV on the other side of the path or once they lose contact with the last node-MAV they dynamically change their type and become a node-MAV.

The pheromone saved within a node-MAV gets increment once ant-MAVs walk by, but it also evaporates over a time and is limited by a maximum number. Once the pheromone is entirely evaporated the state of the MAV changes to ant-MAV. The MAV then returns back to the base node. From there

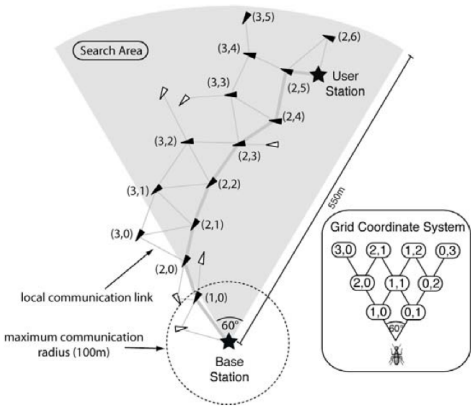


Figure 6. A simulation of 20 MAVs establishing a link from a base station to a user station. Node-MAVs (black triangles) build a grid through which the ant-MAVs navigate using pheromone based rules. Ant-MAVs virtually deposit pheromone through local communication with the node-MAVs. The image is taken from [13].

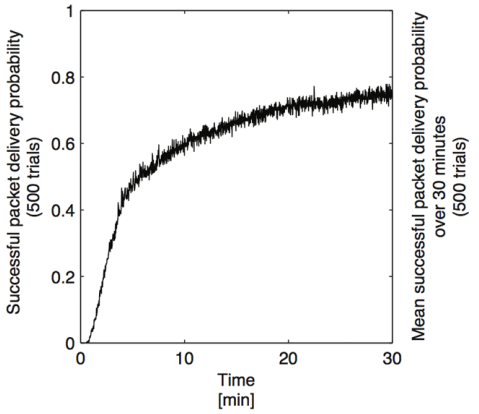


Figure 7. The mean probability of succesfully delivering a packet increases till the end of the trial. This can be explained by the self optimizing mechanism of the MAVs: they are attracted to useful positions within the communication network. The image is taken from [12].

it is redeployed again, meaning it selects a destination node based on the pheromone information in the base node. Then it navigates through the grid to the destination node, checks the pheromone information and selects a new destination node. While navigating towards the destination node the communication link with the last reference node will eventually break. The ant-MAV then waits for contact with the destination node. If there is no contact with the destination node after a certain time the ant-MAV will assume that is doesn't exist and change its type from ant-MAV to node-MAV. The coordinates of the new node-MAV are corresponding to the aimed destination node. The new node-MAV will also have an initial amount of pheromone.

The branch which an ant-MAV chooses to fly on is probabilistically chosen. The equations for the probability were determined on base of the natural model described in [9]. Following the model ant-MAVs favor branches with a higher pheromone amount, but there is also a parameter ensuring that unexplored directions are favored. Initially the choices are made randomly.

Through taking care of unexplored directions it is ensured, that the MAVs will eventually search the whole area. The goal of the grid navigation is not to occupy the whole search area with a grid of MAVs. The goal rather is to dynamically establish a grid in one direction and then slowly adapt it to other directions.

The problem of the swarm splitting up was taken care of by constraining the type-change of ant-MAVs to node-MAVs. Their neighbor node has to be a direct neighbor.

### *C. Establishing a connection*

Once a user station is found the MAVs have to build and maintain a communication link from a base station to the user station. A pheromone trail is used to signal the path from the base station to the user station.

Once a rescuer is found and a link is established the network structure is optimized and stabilized. MAVs are attracted to useful positions in the network using the virtually deposited pheromone trails. This is realized by the node-MAVs being able to detect if they are on a communication path which requires the fewest packet hops from the base station to the user station, the least-hop-route. If they are positioned along such a route they will increase the pheromone on their node. The pheromone in node-MAVs which are not on a route to the user station or further away from the user station will eventually evaporate. They will then navigate back to the base node and redeploy to a new destination node. This ensures that all MAVs will eventually be aligned on a path towards the user station. This way they also form redundant pathways from the base station to the user station. This results in a self-optimizing network as can be seen in Fig. 7. The probability of a successful packet delivery steadily rises.

As shown in Fig. 6 an ideal grid would consist of node-MAVs who position themselves in a grid of Y-branches. The distance between two nodes on a branch should be approximately equal to the communication range between the two MAVs ($\approx$ 100 m) and the junction angle is approximately 60°. This angle was chosen because it maximizes the area coverage of the grid while at the same time generating redundant communication pathways.

MAVs are able to position this way because they are each assigned coordinates $(i, j)$ in the grid. Here i and j are the number of branches the MAV needs to follow to reach a certain position. This is possible because the MAVs contain a magnetic compass and are launched from a fixed point, the base station, the root-node of the grid. Each node can be addressed by labeling them this way.

The problem of colliding MAVs has been taken care of by implementing altitude differing. The MAVs adapt their altitude due to surrounding MAVs.

### *D. Performance*

The qualitative behavior of the swarm was evaluated using a 3D simulation. In 500 trials the swarm was able to find a randomly positioned user station in 91% of the cases. It was found that the performance of an established connection increases until the end of the trial. This can be seen in Fig. 7.

Only 2.6% of the 7.500 deployed MAVs collided with another MAV.

Through the trials the robustness of the network has been extensively tested. It has been found that even swarms with only 5 MAVs are able to find nearly 50% of the users over 500 trials and maintain a connection with them. Even with up to 98% packet delivery success. The performance of the SMAVNET however increases once more MAVs are deployed, which is due to the fact that an area can be searched faster. Also small swarms have difficulties reaching user stations positioned far away.

An animation of the simulation can be seen in the video material [14].

## V. Further Research

### A. Reality

The project in [15] aims on building swarm robots to validate the research previously described. Those are currently developed only in simulation and must be tested in real-life scenarios. The fixed-wing MAVs are built out of Expanded Polypropylene (EPP). They are light-weight (420 g) and have an 80 cm wingspan. They are equipped with a motor on the back and two control surfaces for controlling height and left/right movements. The MAVs are equipped with a micro-controller which gets input from 3 sensors: a gyroscope and two pressure sensors. The communication-based controllers are realized through an ARM based one chip computer running GNU/Linux. Communication with other MAVs is done using an USB WiFi dongle. The dongles were configured to use the 802.11n standard and to transmit in the 5 GHz band. This frequency was chosen because of less estimated interference with other devices, opposed to 2.4 GHz. The dongles were set in ad-hoc mode and have a communication range of about 500 m.

The flight patterns for the MAVs are derived from theses inputs, the controllers output a desired speed, turn rate or altitude which is adapted by the MAV. Under trials in real-life conditions it was found, that abrupt relative displacements between the MAV and a rescuer are a problem. Both approaches presented had the implicit assumption that the MAVs move constantly and freely. Windy conditions however might translate the MAV or even the whole swarm to a completely different location. Since the MAVs don't have any knowledge on a global state this will get quite a challenge to solve. One proposed solution is to frequently replace the MAVs by launching new ones which take the same position. The simulation results indicate that this largely reduces the drift of a swarm.

In the scenario described a rescuer would use the same WiFi dongle and software as used with the MAVs in order to communicate with them. Other research [12] proposes to use the 802.11b specification because it fits best for the technology most ground users use.

### B. Other applications

There is other research on flying ad-hoc networks in a completely different context: Several projects work on using the idea of flying ad-hoc networks in a commercial context. They aim on establishing a broadband communication link into airplanes. The so called aeronautical ad-hoc networks use airplanes as hops to route packets. There is specific research done in the field of establishing a network over flight corridors like the north-atlantic [16].

Another broad research field are aerial swarm systems, organized through a global positioning system. The research aims on achieving tasks such as environmental monitoring (fire detection, toxic plume tracking), area surveillance or tracking and destroying objects.

There are also commercial applications using a swarm of flying robots equipped with cameras to build a high-resolution photographic map of an area.

## VI. Conclusion

Flying swarms of robots are especially suited for applications in disastrous areas because of their flexibility and easy deployability. Since the swarm is homogenous this allows to easily remove, replace or add individuals. This makes a scalable swarm.

We have shown how different researchers aim on building swarms of MAVs without any knowledge of global positioning. To reach this goal a swarm with emergent behavior has to be created. For this purpose two approaches have been presented. They both assume a scenario where MAVs are started from a base station and search for a user station. Once found a communication link has to be established and maintained.

The first approach presented aims at artificially generating neural network controllers for the MAVs. The second approach presented aims on implementing a static set of rules, originally found in foraging patters of army ants, into controllers for MAVs.

There has not been done any direct qualitative comparison between the two approaches. The simulations conducted can not be directly related since they have slightly different conditions. The search area is chosen slightly different and the simulation was conducted in different dimensions. The ant inspired network was simulated using a 3D simulation whereas the artificial evolution approach was done in 2D space. However, both approaches are able to find more than 90% of the user stations and are able to successfully build and maintain a communication link to them. In the army ant inspired approach the network connectivity is steadily optimized due to the self optimizing nature of the network. This might be a possible indication for a better solution.

Both approaches have been tested through extensive simulation and there is current work done on translating the results into reality. This way algorithms could be validated in reality.

Through the recent Arab Spring revolutions it has become clear how repressive governments use the internet as a way to find and eliminate dissidents. A non-constrained, secure internet access is essential for the security of activists and protestors. One could imagine that the approaches described in this article could also be used to supply an area with an non-constraining, secure uplink to the internet. The approach

of a small, light-weight MAV swarm is ideally suited for this scenario since the whole swarm can be operated by a single person who launches them manually by hand.

## REFERENCES


[1] R. Siegwart and I. R. Nourbakhsh, *Introduction to Autonomous Mobile Robots*. Scituate, MA, USA: Bradford Company, 2004.

[2] E. S. Oh, "Information and communication technology in the service of disaster mitigation and humanitarian relief," in *Communications, 2003. APCC 2003. The 9th Asia-Pacific Conference on*, vol. 2, sept. 2003, pp. 730 – 733 Vol.2.

[3] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *SIGGRAPH Comput. Graph.*, vol. 21, pp. 25–34, August 1987. [Online]. Available: http://doi.acm.org/10.1145/37402.37406

[4] "Evolved swarming without positioning information: an application in aerial communication relay," *Auton. Robots*, vol. 26, pp. 21–32, January 2009. [Online]. Available: http://dx.doi.org/10.1007/s10514-008-9104-9

[5] S. Hauert, J.-C. Zufferey, and D. Floreano, "Reverse-engineering of Artificially Evolved Controllers for Swarms of Robots," in *Proceedings of the IEEE Congress on Evolutionary Computation*, 2009, pp. 55–61.

[6] K. Støy, "Using situated communication in distributed autonomous mobile robotics," in *In Proceedings of the 7th Scandinavian conf. on Artificial Intelligence*, 2001.

[7] S. Nolfi and D. Floreano, *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. MIT Press, 2001. [Online]. Available: http://mitpress.mit.edu/catalog/item/default.asp?type=2\&\#38;tid=3684

[8] S. Schimmel, "Bio-inspired networking," in *Proceedings of the Seminar "Research Trends in Media Informatics"*. Universität Ulm. Fakultät für Ingenieurwissenschaften und Informatik, 2010, pp. 583–607. [Online]. Available: http://vts.uni-ulm.de/doc.asp?id=7214

[9] J. L. Deneubourg, S. Goss, N. Franks, and J. M. Pasteels, "The blind leading the blind: Modeling chemically mediated army ant raid patterns," *Journal of Insect Behavior*, vol. 2, no. 5, pp. 719–725, 1989. [Online]. Available: http://www.springerlink.com/index/10.1007/BF01065789

[10] N. R. Franks, N. Gomez, S. Goss, and J. L. Deneubourg, "The blind leading the blind in army ant raid patterns: Testing a model of self-organization (hymenoptera: Formicidae)," *Journal of Insect Behavior*, vol. 4, pp. 583–607, 1991, 10.1007/BF01048072. [Online]. Available: http://dx.doi.org/10.1007/BF01048072

[11] R. V. Solé, E. Bonabeau, J. Delgado, P. Fernández, and J. Marín, "Pattern formation and optimization in army ant raids," Santa Fe Institute, Working Papers, 1999. [Online]. Available: http://econpapers.repec.org/RePEc:wop:safiwp:99-10-074

[12] S. Hauert, L. Winkler, J.-C. Zufferey, and D. Floreano, "Ant-based swarming with positionless micro air vehicles for communication relay," *Swarm Intelligence*, vol. 2, no. 2-4, pp. 167–188, 2008. [Online]. Available: http://www.springerlink.com/index/10.1007/s11721-008-0013-5

[13] ——, "Pheromone-based Swarming for Position-less MAVs," 2007. [Online]. Available: http://lis.epfl.ch/smavs

[14] Ant-inspired control for swarms of flying robots. [Online]. Available: http://youtu.be/5DbWVEQ-JzA?t=35s

[15] S. Hauert, S. Leven, J.-C. Zufferey, and D. Floreano, "Communication-based Swarming for Flying Robots," in *Proceedings of the Workshop on Network Science and Systems Issues in Multi-Robot Autonomy, IEEE International Conference on Robotics and Automation*, 2010.

[16] D. Medina, F. Hoffmann, S. Ayaz, and C.-H. Rokitansky, "Feasibility of an aeronautical mobile ad hoc network over the north atlantic corridor," in *SECON*. IEEE, 2008, pp. 109–116.