Information Hiding im Smart-Grid

Marco Neumann

8. September 2014

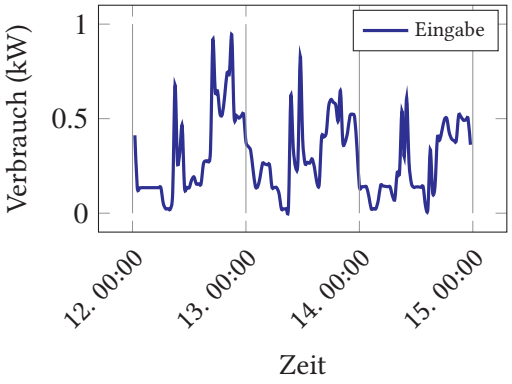


Abbildung 1: Aufzeichnung über 3 Tage im 30-Minuten-Takt

1. Motivation

Eines der großen Probleme unserer Gesellschaft und Ökonomie ist die Versorgung mit Energie. Im Zeichen des Klimawandels muss diese Versorgung in einer sauberen und flexiblen Art und Weise gestaltet werden. Dies ist nur möglich, wenn möglichst alle verfügbaren regenerativen Quellen genutzt werden. Da diese Quellen jedoch nicht kontinuierlich Energie bereitstellen, sind auch für die Verbraucher Änderungen zu erwarten. So wird Strom in Zukunft einen stündlich wechselnden Preis haben und Konsumenten auf den Markt schneller reagieren müssen. Dies ist nur möglich, wenn der Verbrauch entsprechend fein granuliert bestimmt und ausgewertet werden kann. Dazu wurden auf mehreren politischen Ebenen Bestimmungen zum Einbau von sogenannten Smart-Metern eingeführt [Bun13; Eur06]. Diese Stromzähler sind in der Lage, den Verbrauch in bis zu 15-minütigen Abschnitten zu messen. Abbildung 1 zeigt Beispieldaten, welche auch in den weiteren Kapiteln Verwendung finden werden. Sie stammen aus [Ene11].

Neben einer exakten Abrechnung ermöglicht die Erfassung den Verbrauchern auch, ihren Konsum genauer zu kontrollieren und zu optimieren. Die Daten sollen weiterhin Analysten und Planern zur Verfügung stehen, damit diese ein stabiles Netz und optimale und ökonomische sowie umweltverträgliche Erzeugung sicherstellen können.

Diese Weitergabe von Daten birgt jedoch auch viele Risiken. Da die entstandenen Zeitreihen Rückschlüsse auf das persönliche Verhalten der Konsumenten sowie der eingesetzten Geräte in den jeweiligen Haushalten erlauben, ist eine Identifikation der Haushalte möglich und die Persönlichkeitsrechte sind gefährdet. Da die Daten zu Analyse-Zwecken jedoch unbedingt gebraucht werden und die Einführung von Smart-Metern bereits in vollem Gange ist [IDC12a; IDC12b; Eur14], werden Verfahren benötigt, die bestimmte persönliche Informationen aus den Zeitreihen entfernen können. Dieser Vorgang wird in der Fachliteratur Information Hiding genannt.

Im Folgenden werde ich zuerst klassische Ansätze darstellen, die das Problem durch Zugriffsbeschränkungen zu lösen versuchen. Danach werde ich die Grundlagen des Information Hiding und zwei unterschiedliche Ansätze darstellen, gefolgt von einer Analyse dieser Verfahren. Abschließend werde ich einen Ausblick auf zukünftige Forschungen geben und die Auswertungen um ein Plädoyer ergänzen.

2. Bisherige Ansätze

Um den Datenschutz zu verbessern, muss zunächst das zugrundeliegende Problem beschrieben werden. Dieses besteht darin, dass Unbefugte an Informationen gelangen, die nicht für sie bestimmt sind. Deshalb versuchen viele Ansätze, gerade diesen Zugriff zu vermeiden. So wird zum Beispiel in [SK12] versucht, mithilfe von Attribute-Based-Encryption eine Schlüssel-Infrastruktur zu etablieren, die das Zurückziehen von Zugriffsrechten erlaubt. Dies schränkt zwar den Kreis der potentiellen Analysten ein und senkt damit das Missbrauchsrisiko. In Zeiten der NSA-Affäre, international operierenden Unternehmen und zunehmenden Attacken auf Datensilos sollte jedoch jedoch folgende Annahme Grundbestandteil jedes Verfahrens zum Datenschutz sein: Daten, einmal herausgegeben wurden, können potentiell in die falschen Hände geraten. Damit ist ein derartiges Verfahren nicht zielführend.

Ein anderer Ansatz ist der Versuch, spezielle Aggregierungsverfahren zu verwenden [RN10]. Dabei bilden mehrere Haushalte zusammen ein Aggregat und geben dieses an etwaige Analysten weiter. Das vorgestellte Verfahren stellt dabei sicher, dass keiner der Haushalte die exakten Werte seiner Nachbarn weiß und trotzdem ein hinreichend genaues Endergebnis ensteht. Dies erfordert jedoch eine hinreichend große Nachbarschaft, um die individuellen Eigenschaften der Verbraucher im Aggregat zu verbergen. Weiterhin ist ein gewisses Grundvertrauen nötig, da die meisten aggregierte Summen dann nicht funktionieren, wenn man eine hinreichend große Anzahl der Nachbarn unterwandert. Weiterhin ist das Resultat nicht zur börsenbasierten Abrechnung der einzelnen Haushalte geeignet und die benötigte Infrastruktur sowie entsprechende Kompatilitätsprobleme erschweren die Einführung solcher Methoden.

Somit lässt sich zusammenfassend sagen, dass diese Methoden nur unzureichend zum Information Hiding geeignet sind. Es wäre zielführend, wenn die Konsumenten individuell und ohne notwendiges Vertrauen bestimmte Merkmale aus den Daten entfernen könnten. Im Folgenden werde ich zwei Ansätze vorstellen, welche die aufgezeigten Schwächen nicht haben. Es folgt ein Vergleich und ein Ausblick. Fachbegriffe, die zur Erklärung notwendig sind, werden in Anhang A vorgestellt.

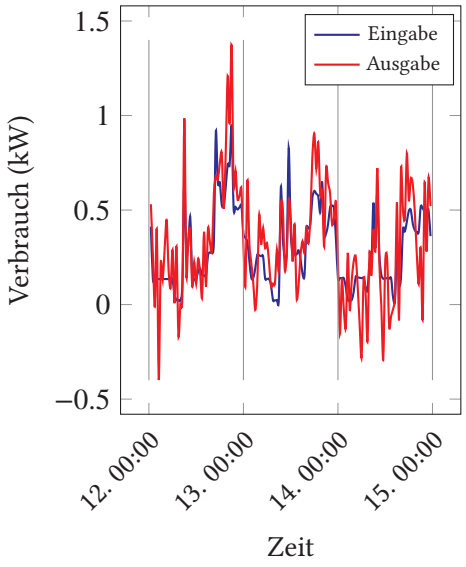
3. Verfahren 1: Verrauschen der Transformierten

In diesem Kapitel werde ich das Verfahren aus [Pap+07] vorstellen

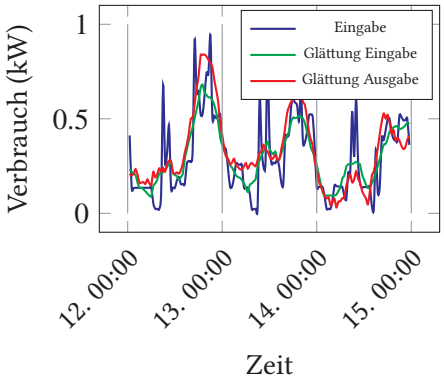
Es ist leicht einzusehen, dass das simple Verrauschen einer Zeitreihe ein schwaches Verfahren des Information Fuzzying ist. Durch Glätten lassen sich leicht Eigenschaften des Ursprungssignals zurückgewinnen. Der Grund dafür ist, dass die aufeinanderfolgenden Werte nicht unabhängig voneinander betrachtet werden dürfen. Es ist wahrscheinlich, dass ein Signalpunkt relativ nahe zu seinen zeitlichen Nachbarn liegt oder im Falle eines Sprunges extrem davon abweicht. Solche Sprünge können beim Ein- und Ausschalten bestimmter Geräte entstehen. Ein Verrauschen des Signals sollte diese zeitliche Beziehung also beachten. Dies kann erreicht werden, indem die ursprüngliche Zeitreihe zuerst in ein geeignetes Modell transformiert wird, man diese Kodierung verrauscht und sie schlussendlich rücktransformiert.

Die zeitlichen Abhängigkeiten im Modell dürfen allerdings nicht zu stark ausgeprägt sein, da sonst eine Anfälligkeit gegen True-Value-Leaks besteht. Bei dieser Form des Angriffs kann aus einem bekannten Wert aus der Eingabe und der kompletten Ausgabe des Verrauschens große Teile der Eingabe wiederhergestellt werden. Bei einfachen Verrauschen ist

Abbildung 2: Einfaches Verrauschen



(a) Einfaches normalverteiltes Verrauschen mit $\sigma = 0.2$



(b) Glätten

dies nicht möglich, da das Rauschsignal für alle Zeitwerte unabhängig ist. Das Gegenteil davor wäre, alle Werte mit dem identischen Delta zu verrauschen. Dies würde bei einem True-Value-Leak die komplette Zeitreihe rekonstruierbar machen. True-Value-Leaks sind ein reales Problem, da zum Beispiel die Nicht-Anwesenheit von Personen durch andere Quellen ermittelt und damit auf einen sehr niedrigen Stromverbrauch geschlossen werden kann.

Aus diesem Grund wird ein Parameter $\sigma \in \mathbb{R}_+$ eingeführt, welcher bestimmt, wie stark die der zeitliche Zusammenhang des Rauschens ist und wie schwer es ist, das Rauschen durch Glättung wieder zu entfernen. Je größer σ ist, desto gröber ist das Rauschen und damit unanfälliger für Glättung. Im Gegenzug wird dadurch die Gefahr bei einem True-Value-Leak erhöht. Diese Konstruktion fundiert auf der Art der Angriffe, die die Autoren hier zugrunde legen. Sie gehen davon aus, dass die bei einer Haar-Wavelet-Transformation errechneten Detailkoeffizienten durch Rauschen erzeugt wurde, wenn sie hinreichend klein sind und anschließend zur Filterung entfernt werden können. Da die Veröffentlichung jedoch von sehr fein aufgelösten Daten mit hinreichend glattem Verlauf ausgeht, was bei den Smart-Meter-Daten nicht der Fall ist (siehe Abbildung 1), möchte auf hier nicht weiter auf diese Art des Angriffes eingehen.

Das Verrauschen erfolgt nun wie folgt:

1. Transformation $m = \alpha(f)$
2. Erzeugung eines Korrelierten Rauschsignals $\gamma = R(m, \sigma)$
3. Rücktransformieren des Rauschsignals $\delta = \alpha^{-1}(\gamma)$
4. Verrauschen durch Addition $\tilde{f} = f + \delta$

Im Folgenden stelle ich zwei geeignete Transformationen vor.

3.1. Fouriertransformation

Ein Modell, das sich hierfür anbietet, ist das Frequenzspektrum. Es repräsentiert die Zusammenhänge aller Datenpunkte auf Zeitebene. Die entsprechende Überführung kann mittels diskreter Fouriertransformation erfolgen. Abbildung 5a zeigt die Transformation des Bespieldatensatzes. Anschließend wird ein entsprechendes Rauschsignal erzeugt. Die Rauschstärke ist dabei vom Spektrum abhängig und richtet sich nach der Amplitude. Weiterhin werden Frequenz-Peaks mit niedriger Amplitude entfernt, um fein granulierte Merkmale aus dem Signal zu entfernen. Die Schwelle ist dabei das bereits behandelte σ. Der Peak bei 0 wird ignoriert, da er lediglich einen konstanten Summanden darstellt. Es handelt sich also um eine Kombination aus Information Fuzzying und Information Removal. Abbildung 5b zeigt das Resultat mit $\sigma = 3$. Es wird deutlich, dass die Ausgabe der Eingabe ähnelt und vor allem bei starken Peaks kaum verändert wird.

Das Frequenzspektrum der Fouriertransformation ist global, also auf die komplette Zeitreihe bezogen. Ein sich dadurch ergebender Vorteil ist die mögliche Darstellung aller periodischen Signale mit beliebigen Frequenzen. Es existieren aber auch mehrere problematische Konsequenzen. Zum einen muss es nach Hinzufügen neuer Werte komplett neu berechnet werden. Zum anderen haben die nahezu rechteckigen Änderungen im Stromverbrauch Auswirkungen auf eine breite Masse von Punkten im Spektrum. Da das Verrauschen aller Frequenz-Punkte aber unabhängig voneinander geschieht, ist es schwierig, derartige Schwankungen geeignet zu verschleiern.

3.2. Wavelettransformation

Ein anderen Ansatz zur Transformation bietet die Wavelettransformation im Allgemeinen und das Haar-Wavelet im speziellen. Die weitere Betrachtung wird sich ausschließlich auf dieses eine Wavelet konzentrieren. Der interessierte Leser findet ausreichend Literatur zur Wavelet-Theorie.

Die Haar-Wavelettransformation basiert im Grunde auf einem perfekten High- und Lowpass-Filter. Das Eingangssignal wird geteilt. Das niedrigfrequente Signal hat nur die halbe Samplingrate verglichen zum Eingangssignal. Das hochfrequente Signal kann mit Bezug auf das niedrigfrequente ebenfalls herab gesamplet werden. Da dieser Ansatz jedoch schwer verständlich ist, werde ich einen andere Variante vorstellen. Wie in Abb. 6 gezeigt, kann das Signal wie folgt transformiert werden:

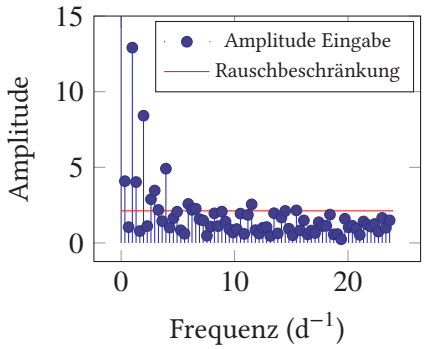
1. Baumartige Durchschnittsbildung:
   Dabei wird jeweils der Durchschnitt von 2 Signalpunkten gebildet. Ist das Signallänge keine Potenz von 2, wird der Signalpunkt am Ende ohne Durchschnittsbildung übernommen.

2. Berechnung der Deltas:
   Nach der Durchschnittsbildung wird der Baum von der Wurzel aus rekonstruiert, indem man den Unterschied vom Vaterknoten zum linken (aka zeittechnisch zeitigerem) Kind berechnet. Für das andere Kind ist der Unterschied einfach das Negative, da der Vaterknoten der Durchschnitt beider Kindsknoten ist.

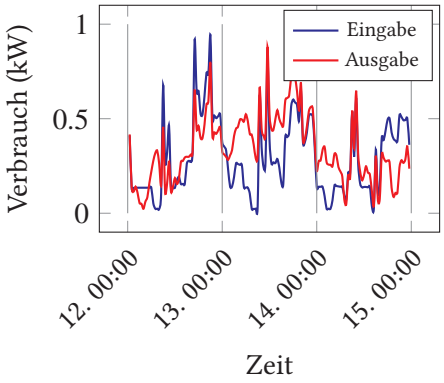
3. Serialisierung des Delta-Baums:
   Das transformierte Signal ist eine Travesierung des Delta-Baums.

Das ursprüngliche Signal kann aus dem Delta-Baum durch Travesierung von der Wur-

Abbildung 4: Fouriertransformation



(a) Fouriertransformation des Beispiel-Datensatzes. Cut bei $\sigma = 3$. Peak bei 0 irrelevant.



(b) Ergebnis der Fourier Methode

Abbildung 6: Haar Wavelets

Original Signal	2	1	3.5	3.5	3	2	0.5	–
Durschnitt 2	1.5		3.5		2.5		0.5	
Durschnitt 4		2.5				1.5		
Durschnitt 8				2				

(a) Schritt 1: Durchschnitte berechnen

Anker				2				
Delta Stufe 0				0.5				
Delta Stufe 1		−1				1		
Delta Stufe 2	0.5		0		0.5		–	
Original Signal	2	1	3.5	3.5	3	2	0.5	–

(b) Schritt 2: Delta ermitteln

Transformiertes Signal	2	0.5	−1	1	0.5	0	0.5	–

(c) Schritt 3: Ausgabe

zel zu den Blattknoten gewonnen werden. Dafür wird die Summe aus der Wurzel und den Deltas gebildet. Für rechte Kindsknoten wird das negative Delta verwendet. Abbildung 9a zeigt die Transformation des Musters.

Das Verrauschen ist ähnlich zum Verfahren der Fouriertransformierten. Der Anker wird nicht beachtet und einfach genullt und die Detailkoeffizienten werden einzeln verrauscht. Abbildung 9b zeigt das Ergebnis der Haartransformationsmethode.

Im Gegensatz zur Fouriertransformation haben Peaks in der Haar"=Wavelettransformierten nur Auswirkungen in einem beschränkten Bereich mit abnehmender Stärke proportional zur Entfernung in der Baumstruktur. Dies behebt die angesprochenen Probleme teilweise.[1] Allerdings ergibt sich durch die Lokalität auch ein neues Problem. Periodische Signale können Einfluss auf große Teile der Transformierten haben. Da das Verrauschen der Elemente der Transformierten unabhängig voneinander erfolgt, lassen sie sich nach dem Information Fuzzying und Removal leicht wiederherstellen. So kann dem Angreifer zum Beispiel bekannt sein, dass ein Kühlschrank periodische Kühlphasen hat, was zu einem periodischen Stromverbrauch führt. Da die einzelnen Perioden jedoch unabhängig voneinander verrauscht werden, können diesen unabhängigen Rauschelemente zumindest teilweise herausgerechnet werden. Es handelt sich hierbei nicht direkt um einen True-Value-Leak, da nicht die echten Werte selbst, sondern ihre relativen Zusammenhänge bekannt sind.

[1] Nicht vollständig, wie hier visualisiert http://i08fs1.atis-stud.uni-karlsruhe.de/~s_mneuma/visual/haar_perturbation/

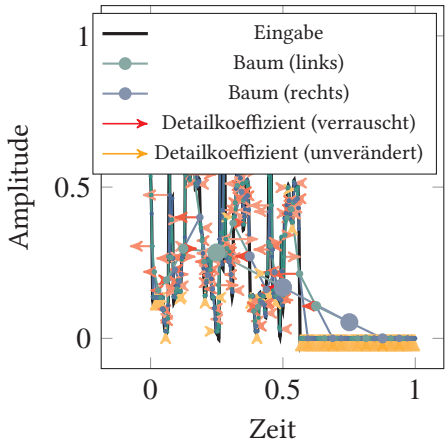
4. Verfahren 2: Markov Ketten und Batterie-Puffer

Während des bereits vorgestellte Verfahren versucht, Informationen in der Messung selbst zu verstecken, setzt [Kal+11] auf ein anderes Modell. Hierfür wird angenommen, dass Geräte entweder ein- oder ausgeschaltet sein können und die Menge der interessanten Geräte bekannt sind. Neben interessanten Geräten gibt es auch solche, deren Zustand ohne größere Bedenken geleakt werden können. Ein Beispiel dafür ist der Kühlschrank. Dieses Modell entspricht auch hauptsächlich den Daten, die man aus den Verbrauchsdaten gewinnen möchte, um daraus das Verhalten des Verbrauchers zu analysieren. Das sich daraus ergebende Modell ist das Hidden Markov Model. Normalerweise ist es möglich, durch den Stromverbrauch die internen Zustand der Markov Kette relativ gut zu schätzen. Was gut bedeutet, hängt stark von den bekannten Geräte, der Stabilität ihres Verbrauchs und dem Unterschied der von ihnen verursachten Last ab. [Kal+11] versucht nun, durch geeignete Batterie-Puffer und einen intelligenten Algorithmus den Verbrauch pro Gerät möglichst konstant zu halten und damit die Zustandsschätzung der Markovkette zu erschweren.

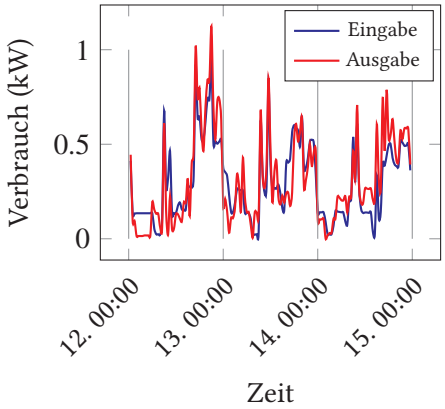
4.1. Grundlegendes Verfahren

Um den Verbrauch der Geräte konstant zu halten, müssen auftretende Schwankungen durch einen Batteriepuffer ausgeglichen werden. Für die verwendete Batterie sind Kapazität sowie maximaler Lade- und Entladestrom bekannt. Diese sind in der Regel nicht identisch. Außerdem wird die Kapazität in Partionen unterteilt, wobei jedes schützenswerte Gerät eine eigene erhält. Die Größe der Partitionen richtet sich dabei nach dem Verbrauch und der erwarteten Schwankung des Geräts, aber auch nach dem gewünschten Schutzlevel. So kann es bei Fehl-

Abbildung 8: Haar-Methode



(a) Haar-Transformation des Beispiel"=Datensatzes



(b) Ergebnis der Haar Methode

einschätzung oder langen Ein- oder Ausperioden dazu kommen, dass der Zustand des Gerätes geleakt werden muss. Dies ist bei besonders prisanten Geräten zu vermeiden. Zusätzlich wird jedes Gerät mit einem eigenen Strommesser ausgestattet, um den aktuellen Zustand ermitteln zu können. Desweiteren werden alle Messgeräte mit Gateways verbunden und diese an ein Steuergerät für die Batterie angeschlossen.

Es wird angenommen, dass der Verbrauch über feste Zeitabstände akkumuliert und gemeldet wird. In diesen Intervallen wird nur der durchschnittliche Verbrauch berücksichtigt und eventuell auftretende Peaks außer Acht gelassen. Dies entspricht dem Verhalten der meisten Smart-Meter. Im Gegenteil zu dem Verfahren aus Abschnitt 3 wird die Änderung am Stromverbrauch real und vor der Messung durch das Smart-Meter vorgenommen. Ein Eingriff in die Messung oder Meldung der Daten ist also nicht notwendig.

Pro Zeiteinheit werden nun folgende Schritte durchgeführt:

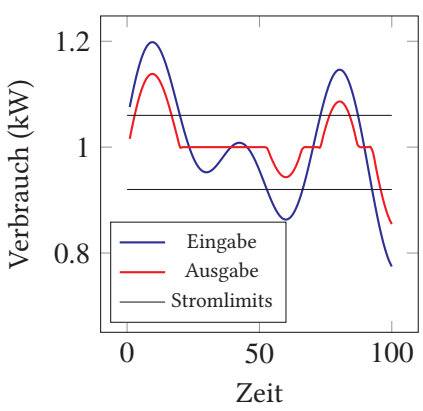
1. Schätze Verbrauch pro Gerät mittels Hidden-Markov-Models
2. Beschränke akkumulierten Batteriestrom
3. Beschränke Partionsladung

Für die Schätzung durch das Hidden-Markov-Model möchte ich auf [Kal+11] verweisen.

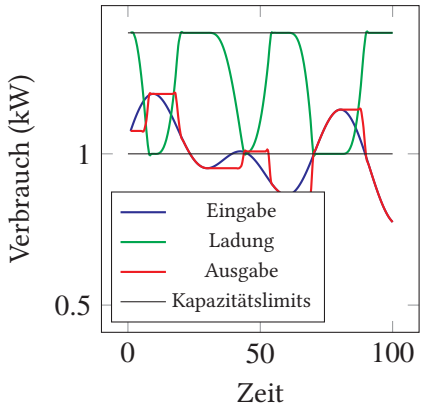
4.2. Beschränkung: Gesamt-Strom

Da die Batterie nur einen bestimmten Lade- und Entladestrom zur Verfügung stellt, muss der Ausgleichsstrom unter Umständen beschränkt werden. Dieser besteht aus der Summe alle ladenden und entladenden Geräte, was häufig eine Gleichgewichtssituation zur Folge hat. Im Falle der Beschränkung wird der Strom relativ auf alle Lader bzw. Entlader aufgeteilt. Große Verbraucher bekommen also einen entsprechend größeren Anteil. Abbildung 11a zeigt ein solche Situation. Im Falle der Überschreitung der Limits wird die Differenz zum Maximum geleakt und alle Geräte sind gleichermaßen betroffen.

Abbildung 10: Batterie Limits



(a) Strom Limits



(b) Kapazität Limits

4.3. Beschränkung: Partitions-Ladung

Im Falle großer Schwankungen oder langer Ein- bzw. Ausphasen von Geräten kann es dazu kommen, dass die entsprechende Partition voll oder komplett leer wird. In diesem Fall wird die restliche Differenz verbraucht und danach eine Zustandsänderung geleakt. Abbildung 11b zeigt ein derartiges Beispiel. Man beachte, dass die Ladung einer Partition keinen Einfluss auf den Zustand einer anderen hat und die Partitionsgrößen konstant bleiben. Ein intelligenteres Management ist Aufgabe zukünftiger Forschungen.

4.4. Zusammenfassung

Schlussendlich wird der für jedes Gerät errechnete Ausgleichsstrom addiert und somit die Batterie geladen oder entladen. Abbildung 12 zeigt ein Beispiel mit 2 kW. Da für den Musterdatensatz keine Dekomposition vorliegt, wird der gesamte Haushalt als ein einziges Gerät betrachtet. Es wird sichtbar, dass vereinzelte Leaks auftreten können, da die Batterie nicht hinreichend groß ist.

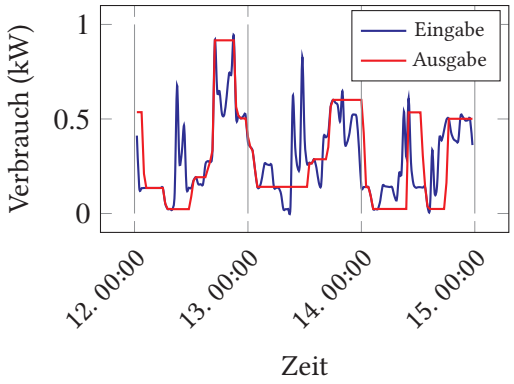


Abbildung 12: Ergebnis der Batterie Methode. Ein Gerät und 2 kW

5. Vergleich

Nachdem nun beide Verfahren vorgestellt worden, möchte auf die Unterschiede und die sich daraus ergebenden Anwendungszwecke eingehen.

5.1. Usability

Die einfache Einrichtung und Benutzbarkeit einer Datenschutzmethode ist von großer Bedeutung. Nur durch diese Faktoren ist eine schnelle und breitflächige Einführung möglich.

Das Transformieren durch geeignetes Verrauschen ist eine rein mathematische Methode. Dies ermöglicht eine Umsetzung auf Softwareebene und damit eine einfache Installation auf dem Smartmeter oder Gateway. Der Rauschparameter kann und soll vom Hersteller vorgegeben oder von der Software geschätzt werden. Dies ist nötig, da dem Verbraucher das nötige Fachwissen fehlt. Da viele Smartmeter ohnehin mit Zusatzmodulen oder Apps werben, ist bereits ein Vertriebsweg vorhanden. Wartung und Nachkonfiguration ist in der Regel nicht nötig und damit eine komfortable Nutzung gewährleistet.

Dem gegenüber steht das Batterie-basierte Verfahren. Hier sind eine Anschaffung und Installation von Batterie, Strommessern für alle schützenswerte Geräte inkl. Gateways und Verkabelung sowie eine Schätzung der Partitionen notwendig. Dies ist ohne Fachwissen und Erfahrung durchaus anspruchsvoll und sollte von Fachpersonal übernommen werden. Der invasive Eingriff in die Infrastruktur des Haushaltes dürfte dabei eine große Hürde darstellen, genauso wie die Hürde zum Kaufen und Einbauen der Technik. Die Hürde sinkt deutlich, wenn bereits eine Batterie oder entsprechende Messtechnik im Haus vorhanden sind. Dies ist zur Zeit beim Durchschnittsverbraucher jedoch nicht der Fall.

Aus aktueller Sicht ist die Usability ist aufgrund der genannten Gründe für das erste Verfahren weitaus besser und sollte bei marktstrategischen Entscheidungen berücksichtigt werden.

5.2. Verwertbarkeit

Der zweite wichtige Vergleichspunkt ist die Verwertbarkeit der produzierten Daten. Eines der Hauptziele der Einführung von Smartmetern ist die Nutzung der Daten durch Dritte. [Eur12] Dafür ist zu bewerten, ob die Resultate auch entsprechend verwertbar sind.

Die von Verfahren 1 produzierten Daten entsprechen nicht mehr dem eigentlichen Stromverbrauch. Sie sind damit für Abrechnungszwecke ungeeignet und können nur vom Versorger oder Forschern zu Analyse- und Planungszwecken verwendet werden. Dazu ist eine von der Abrechnung getrennte Übertragung notwendig. Da dies von den aktuellen Richtlinien und Gesetzen [Eur06; Bun13] nicht vorgesehen ist, wird eine Neuregelung nötig. Eine derartige Änderung auf europäischer Ebene und die damit verbundene Lobbyarbeit dürfte jedoch schwer sein. Sie entspricht nicht den Interessen der Stromunternehmen und fordert zudem ein Verständnis des Verfahrens durch die Politiker. Dies dürfte für die meisten jedoch Neuland sein. Somit bleibt nur eine nachträgliche Bearbeitung von Forschungsdatensätzen, was höchstens von Universitäten und entsprechenden Instituten, wohl aber kaum von Unternehmen zu erwarten ist.

Bei Verfahren 2 hingegen entspricht der gemeldete dem gemessenen Verbrauch, was eine Nutzung zur Abrechnung ermöglicht und mit der aktuellen Gesetzeslage vereinbar ist. Da keine Abweichung zwischen gemessenen und realem Verbrauch existiert, sind die Daten auch für Versorgern verwertbar.

Vom Standpunkt der Verwertbarkeit aus gesehen ist Verfahren 2 damit deutlich im Vorteil.

5.3. Kosten

Ein für die produktive Nutzung wichtiger Punkt sind die Kosten, die bei der Anwendung eines Verfahrens entstehen. Sie bestimmen neben der Usability hauptsächlich, wie Methode und das darauf aufbauende Produkt oder Dienstleistung vom Verbraucher angenommen wird. Dies gilt auch für Produkte, welche die Privatssphäre schützen sollen.

Für das in [Pap+07] vorgestellte Verfahren fallen lediglich Kosten für die Software oder das Erweitungsmodul an. Ausgehend davon,

dass diese Software gekauft und nicht gemietet wird, fallen keine laufenden Kosten an. Die einmaligen Anschaffungskosten dürften für den Anwender erträglich ausfallen.

Die für die Batterie-Methode anfallenden Kosten unterscheiden sich davon erheblich. Für die meisten Nutzer fallen Anschaffungs- und Installationskosten für Messgeräte, Gateways, Steuergerät und Batterie an. Dazu kommen noch laufende Kosten, welche sich aus dem Verlust bei der Speicherung in der Batterie sowie dem Verschleiß ergeben. Auf Grundlage von [Lei14] ergibt sich für die jede kWh Leistung, die in der Batterie zwischengespeichert wird, Kosten von mindestens 10 ct. Dies sind in dem in Abbildung 12 simulierten Beispiel mindestens 100 € pro Jahr. Falls bereits eine Batterie installiert ist oder Strom durch erneuerbare Energien gewonnen wird, kann dieser Preis sinken.

Unter dem Gesichtspunkt der entstehenden Kosten ist das erste Verfahren zu bevorzugen.

5.4. Privacy

Das Schützen der Privatssphäre der Nutzer ist das Hauptziel der vorgestellten Verfahren und dieser Arbeit. Somit ist es wichtig zu betrachten, wie gut diese Aufgabe erfüllt wird.

Die mathematische Transformation erfüllt das selbst gesteckte Ziel des Schutzes gegen True-Value-Leaks und Glättung gut. Allerdings ist es nicht in Lage, den Zeitpunkt von Peaks zu verschieben und starke Peaks zu verdecken. Damit ist es möglich, mithilfe von Hidden-Markov-Modellen den Zustand zu schätzen oder ein Fingerprinting der Konsumenten durchzuführen. Untern Umständen kann auch die Extraktion bestimmter Information wie die Anzahl der Bewohner, das Vorhandensein bestimmter Geräte oder das geschaute Fernsehprogram möglich sein. Gerade letzteres ist möglich, da auf hohen Frequenzen in der Regel kein weiteres Rauschen hinzugefügt wird. Es eignet sich daher zum Verdecken des eigentlichen Verbrauchs, nicht jedoch zum Schutz gegen Informationsgewinnung und Re-Identifizierung. Ein weiteres Problem ist die getrennte Übertragung zu Abrechnungszwecken, vor allem in börsenartigen Umgebungen. Vorfälle der letzten Jahre haben gezeigt, dass einmal übertragene Daten als geleakt zu betrachten sind und damit das eigentliche Ziel verfehlt wird.

Das in [Kal+11] entwickelte Verfahren ist in der Lage, den Zeitpunkt von Ereignissen zu verschieben oder Änderungen im Zustand von Geräten ganz zu verstecken. Allerdings ist es möglich, aus vermeintlich uninteressanten und nicht schützenswerten Geräten wichtige Informationen zu extrahieren. So ändert sich die Verbrauchssignatur eines Kühlschranks mit Einkaufsverhalten, auch wenn dies von den Autoren nicht beachtet wird. Weiterhin ist es schwer, ausreichend große Partitionen für alle Geräte bereitzustellen. Ein Leak kann daraufhin Aufschluss über die Partitionsgrößen und die versteckten Geräte liefern. Weiterhin kann der akkumulierte Stromverbrauch aller Geräte Aussagen über Anwesenheit von Bewohnern sowie ein Fingerprinting ermöglichen. Dies wurde von den Autoren nicht ausgeschlossen.

Gerade bei der Bewertung des Schutzes der Privatssphäre ist es schwer, die beiden Verfahren zu vergleichen, da diese unterschiedliche Definitionen verwenden und dadurch andere Ziele verfolgen. Beide Methoden bieten keinen vollständigen Schutz. Auch wurde keine Angriffe mit nicht-trivialen Methoden durchgeführt.

5.5. Ausblick und zukünftige Entwicklung

Als finaler Punkt bleibt zu betrachten, wie sich die Verfahren und ihre Anwendbarkeit in Zukunft entwickeln werden und ob eine weiter-

führende Forschung lohnenswert ist. Dies beruht zwar vor allem auf Abschätzungen der weiteren Entwicklung, bietet jedoch eine gute Grundlage für Investitions- und Forschungsentscheidungen.

Für alle Verfahren mit rein mathematischer Grund gilt wie schon besprochen, dass sie nicht den aktuellen Regelungen von EU und BRD entsprechen. Es ist nicht zu erwarten, dass sich dies in naher Zukunft ändert. Desweiteren sind für geplante börsenartige Strukturen ohnehin Daten erforderlich, die mit dem realem Verbrauch übereinstimmten.

Methoden, die auf der Änderung mittels Batteriepuffern basieren, haben diese Nachteile nicht. Desweiteren ist ein stark sinkender Batteriepreis zu erwarten. [US 14; Tes14] Zusätzlich können derartige Algorithmen mit Verfahren zum Handel an Strombörsen oder verschiedenen Tageszeittarifen kombiniert werden. Damit dabei kein Leaking von Daten ensteht, ist weitere Forschungsarbeit nötig. Auch die Kombination mit eigener Stromerzeugung durch erneuerbare Energien ist möglich.

Für zukünftige Entwicklungen, Forschungen und Investitionen sind damit Verfahren, welche auf Batterie-Puffern basieren, deutlich im Vorteil. Es ist jedoch zu beachten, dass sich die Rechtslage jederzeit ändern kann und sich die Einführung von börsenbasiertem Ein- und Verkauf stark verzögern kann.

5.6. Fazit

Die Entscheidung über die Implementierung, Installation und Einsatz der Verfahren hängt stark von den gewünschten Eigenschaften, dem Vorhanden Budget, der angestrebten Privacy und langfristigen Investition ab. Beide Verfahren haben individuelle Vor- und Nachteile. Somit steht auf der einen Seite mit dem in [Pap+07] ein sofort einsetzbares und preisgünstiges Modell zur Verfügung, welches sich vor allem zur Gewinnung von Forschungsdatensätzen eignet. Auf der anderen Seite wird in [Kal+11] ein Aufbau präsentiert, welcher gerade aufgrund der Kosten noch etwas Zeit bis zum Einsatz benötigt, mit der Kombinierbarkeit mit anderen Verfahren und der rechtlichen Sicherheit jedoch ein guter Kandidat für zukünftige Forschungen ist.

6. Ausblick

Beide Paper wurden bisher nicht gegen moderne Angriffsmethoden getestet. Ein wichtiger offener Punkt ist somit das Testen gegen Verfahren aus [BSS13], [Buc+13], [Kim+11] und [Wij+14]. Desweiteren können Batterie-Verfahren mit Speichern für erneuerbare Energien kombiniert werden [TGP12] und weiter Alternativen wie [Raj+11] in den Vergleich einbezogen werden.

Literatur

[BSS13] Christian Beckel, Leyna Sadamori und Silvia Santini. „Automatic Socio-economic Classification of Households Using Electricity Consumption Data“. In: *Proceedings of the Fourth International Conference on Future Energy Systems*. e-Energy '13. Berkeley, California, USA: ACM, 2013, S. 75–86. isbn: 978-1-4503-2052-8. doi: 10.1145/2487166.2487175.

[Buc+13] Erik Buchmann u. a. „Re-identification of Smart Meter data“. In: *Personal and Ubiquitous Computing* 17.4 (2013), S. 653–662.

[Bun13] Bundesrepublik Deutschland. *Energiewirtschaftsgesetz §21c*. 2013.

[Ene11] Commission for Energy Regulation (CER). *Customer behaviour trials findings report (cer11/080a)*. Techn. Ber. 2011.

[Eur06] Europäische Union. *Richtlinie 2006/32/EG über Endenergieeffizienz und Energiedienstleistungen*. 2006.

[Eur12] Europäische Kommission. *European Task Force Tor The Implementation Of Smart Grids Into The European Internal Market*. 2012. url: http://ec.europa.eu/energy/gas_electricity/smartgrids/doc/mission_and_workprogramme.pdf.

[Eur14] Europäische Kommission. *Cost-benefit analyses & state of play of smart metering deployment in the EU-27*. 2014. url: http://eur-lex.europa.eu/legal-content/EN/NOT/?uri=SWD:2014:189:FIN.

[IDC12a] IDC Corporate USA. *Global smart meter shipment forecast from 2010 to 2016 (in million units)*. Statista, Inc., 2012. url: http://www.statista.com/statistics/241989/global-smart-meter-shipment-forecast/.

[IDC12b] IDC Corporate USA. *Smart meter shipments in the Americas from 2010 to 2016 (in million units)*. Statista, Inc., 2012. url: http://www.statista.com/statistics/241993/smart-meter-shipment-forecast-in-america/.

[Kal+11] Georgios Kalogridis u. a. „ElecPrivacy: Evaluating the Privacy Protection of Electricity Management Algorithms“. In: *IEEE Trans. Smart Grid* 2.4 (2011), S. 750–758.

[Kim+11] Hyungsul Kim u. a. „Unsupervised Disaggregation of Low Frequency Power Measurements“. In: *SDM*. SIAM / Omnipress, 2011, S. 747–758. isbn: 978-0-898719-92-5.

[Lei14] Leipziger Institut für Engergie GmbH. *Wirtschaftlichkeit Batteriespeicher*. 2014. url: http://www.ie-leipzig.com/010-dateien/referenzen/pdf/ie_2014-01-29_wirtschaftlichkeit-batteriespeicher_kurzexpertise.pdf.

[Pap+07] Spiros Papadimitriou u. a. „Time Series Compressibility and Privacy“. In: *VLDB*. Hrsg. von Christoph Koch u. a. ACM, 2007, S. 459–470. isbn: 978-1-59593-649-3.

[Raj+11] S. Raj Rajagopalan u. a. „Smart meter privacy: A utility-privacy framework“. In: *SmartGridComm*. IEEE, 2011, S. 190–195. isbn: 9781457717048.

[RN10] Vibhor Rastogi und Suman Nath. „Differentially private aggregation of distributed time-series with transformation and encryption“. In: *SIGMOD Conference*. Hrsg. von Ahmed K. Elmagarmid und Divyakant Agrawal. ACM, 2010, S. 735–746. isbn: 978-1-4503-0032-2.

[SK12] Jens Strüker und Florian Kerschbaum. „From a Barrier to a Bridge: Data-Privacy in Deregulated Smart Grids“. In: *ICIS*. Association for Information Systems, 2012.

[Tes14] Tesla Motors, Inc. *Gigfactory Presentation*. 2014. url: http://www.teslamotors.com/sites/default/files/blog_attachments/gigafactory.pdf.

[TGP12] Onur Tan, Deniz Gündüz und H. Vincent Poor. „Smart meter privacy in the presence of energy harvesting and storage devices“. In: *SmartGridComm*. IEEE, 2012, S. 664–669. isbn: 978-1-4673-0910-3.

[US 14] US Geological Survey. *Countries with the largest lithium reserves worldwide as of 2013 (in metric tons)*. Statista, Inc., 2014. url: http://www.statista.com/statistics/268790/countries-with-the-largest-lithium-reserves-worldwide/.

[Wij+14] Tri Kurniawan Wijaya u. a. „Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data“. In: *SIAM International Conference on Data Mining (SDM14)*. Philadelphia, Pennsylvania, USA, 2014.

A. Theorie des Information Hidings

Für den interessierten Leser möchte ich klären, was mathematisch unter Information Hiding zu verstehen ist. Die Definitionen entsprechen meinem Verständnis zum Thema und können von anderen Veröffentlichungen zu diesem Thema abweichen. Dazu ist zunächst eine Zeitreihe gegeben:

Definition 1 (kontinuierliche Zeitreihe)**.** Eine kontinuierliche Zeitreihe $f : \mathbb{R}_+ \to \mathbb{R}_+$ ist eine Funktion, welche jeden Zeitpunkt $t \in \mathbb{R}_+$ auf einen positiven Wert $x \in \mathbb{R}_+$ abbildet.

Es sei angemerkt, dass diese Definition nur im Kontext des Energieverbrauchs sinnvoll ist und die Erzeugung nicht mit einbezieht. Außerdem kann die Beschränkung auf positive Zeitpunkte in manchen anderen Arbeiten nicht zielführend sein, wenn zum Beispiel Verläufe relativ zu einem Zeitpunkt untersucht werden sollen. Für diese Arbeit ist diese Einschränkung jedoch hilfreich. Da Zeitreihen aus technischen Gründen mit einer bestimmten Abtastrate gemessen werden müssen, werden ich im folgenden diskrete Zeitreihen verwenden. Desweiteren sind Zeitreihen in Wirklichkeit endlich. In dieser Ausarbeitung gehe ich ebenfalls davon aus, dass in diesem endlichen Intervall alle Werte existent sind. Dies führt zur Definition der diskreten endlichen Zeitreihe:

Definition 2 (diskrete endliche Zeitreihe)**.** Eine diskrete endliche Zeitreihe $f : T \to \mathbb{R}_+$ ist eine Funktion, die jeden diskreten Zeitpunkt $t \in T$ aus dem endlichen und abgeschlossenen Zeitraum $T \subset \mathbb{N}$ auf einen positiven Wert $x \in \mathbb{R}_+$ abbildet. Der Raum aller Zeitreihen wird als F bezeichnet.

Da diese Art der Zeitreihe die einzig praktisch relevante ist, werde ich im Folgenden die Artenbezeichnung weglassen und nur von Zeitreihen sprechen. Nun sind diese Zeitreihen nur selten von direktem Nutzen. Aus ihnen lassen sich keine Informationen wie die verwendeten Geräte oder der Durchschnittsverbrauch direkt ablesen. Sie stellen lediglich den Verbrauch zu den bestimmten Zeitpunkten dar. Die für Analysten interessanten Informationen werden erst in bestimmten Modellen sichtbar:

Definition 3 (Modell)**.** Ein Modell M ist eine Darstellung für Zeitreihen F in einer Art und Weise, dass sich gewünschte Informationen daraus ableiten lassen. Die Elemente aus M werden Kodierungen genannt.

Um eine Zeitreihe in eine Kodierungen zu überführen, muss sie in das Modell abgebildet werden. Neben der Zeitreihe selbst können dafür noch weitere Informationen wie Jahreszeit, Ort der Messung oder Strompreise genutzt werden:

Definition 4 (Transformation)**.** Eine Transformation $\alpha : F \times I \to M$ ist eine Abbildung einer Zeitreihe $f \in F$ und Zusatzinformationen $i \in I$ in eine Kodierung eines Modells $m \in M$.

So ist zum Beispiel das Integral bzw. die Summe über eine Zeitreihe die Transformation zum Modell des Gesamtverbrauchs. Eine Kodierung des Gesamtverbrauchs ist sinnvollerweise eine positive reelle Zahl. Ein anderes Beispiel wäre die Transformation in einen binäre Zeitreihe der Form $b : T \to \{0, 1\}$, welche anzeigt, ob mindestens eine Person im gemessen Haushalt anwesend ist.

Da jetzt alle Grundlagen gelegt wurden, möchte ich den Begriff des Information Removals klären. Dabei soll das Modell so in sich selbst abgebildet werden, dass dabei bestimmte Informationen innerhalb dieses Modells verloren gehen. So kann zum Beispiel der Gesamtverbrauch gerundet werden. Mathematisch korrekt sieht die Definition so aus:

Definition 5 (Information Removal)**.** Information Removal bezeichnet eine Abbildung $z_r : M \to N \subset M$ eines Modells M in sich selbst, dass dabei Informationen verlorengehen. Der Informationsverlust kann dabei durch die Differenz $M \setminus N$ dargestellt werden.

Nach dieser Definition ist das Runden des Gesamtverbrauchs die Abbildung $z_r : \mathbb{R}_+ \to \mathbb{N} \subset \mathbb{R}_+$ mit $z_r(x) = \lceil x \rceil$.

Leider ist das Entfernen von Informationen nicht immer anwendbar. So soll für die Analysten weiterhin brauchbares Material zur Verfügung stehen. Außerdem kann das Entfernen von Informationen selbst eine Informationen sein. So kann das Fehlen von Daten über bestimmte Geräteklassen oder Zeiträume Rückschlüsse darauf zulassen, dass es sich dabei um sensible Informationen handelt. Deshalb müsste die Informationen Removal Abbildung für mehrere Haushalte identisch oder hinreichend ähnlich sein. Dies erschwert das Finden einer geeigneten Abbildung noch weiter. Deshalb werde ich auf eine Alternative zurückgreifen, die auf dem Verwischen von Informationen basiert. Dieses Verwischen wird Information Fuzzying genannt und bezeichnet das Ändern der Kodierung, so dass die Informationen geringfügig anders sind und sich Rückschlüsse nur noch mit einer bestimmten Wahrscheinlichkeit ergeben. Die formale Definition lautet wie folgt:

Definition 6 (Information Fuzzying)**.** Information Fuzzying $z_f : M \to M$ ist eine Abbildung von einer Kodierung $m \in M$ auf eine andere Kodierung $n \in M$, wobei für eine Metrik $d : M \times M \to \mathbb{R}_+$ und ein geeignetes $\epsilon \in \mathbb{R}_+$ gelten muss: $E[d(m, n)] \leq \epsilon$. $E[\cdot]$ beschreibe hier den Erwartungswert.

Das in der Definition genannte $\epsilon \in \mathbb{R}_+$ bestimmt die Stärke des Fuzzying. Für das Beispiel des Gesamtverbrauchs mit der normalen Abstandsmetrik $|a - b|$ wäre ein Information wie folgt möglich: $z_f(x) = n(x, \sigma)$, wobei $n(\mu, \sigma)$ einen zufälligen Punkt nach Normalverteilung mit Erwartungswert μ und Varianz σ^2 beschreibt.

Die zwei vorgestellten Methoden zum Entfernen bzw. Verwischen von Informationen können unter dem Begriff Information Hiding zusammengefasst werden. Information Fuzzying stellt insofern einen Verlust von Informationen dar, dass nicht mehr mit hinreichender Sicherheit gesagt werden kann, dass die schlussendliche veröffentlichte Kodierung mit dem Original übereinstimmt. Der Vorteil eines guten Information Fuzzying ist, dass nicht sofort ersichtlich ist, ob und wie viele Informationen versteckt werden sollen.

An dieser Stelle sei zusätzlich angemerkt, dass durch das Überführen eines Modells in ein weiteres zusammen mit Zusatzinformationen das Information Hiding zum Teil übergangen werden kann. So können bestimmte Arten von Rauschen geglättet werden und damit

ursprüngliche versteckte Informationen wiederhergestellt werden. Diese weiteren Transformationen stellen Angriffsvektoren gegen Verfahren des Information Hidings dar. In der Regel lassen sich nur bestimmte Klassen von Angriffsvektoren ausschließen. Diese Klassen hängen stark mit dem Zielmodell des Angriffs zusammen. So können durch bestimmte Arten des Rauschens meist bestimmte Arten der Glättung ausgeschlossen werden. Andere Transformationen können jedoch digitale Fingerabdrücke oder sogar Teile der ursprünglichen Kodierung wiederherstellen. Resistenz gegen Angriffsvektoren soll in dieser Arbeit nur am Rand behandelt werden, da dies zusammen mit entsprechenden mathematischen Beweisen den Rahmen sprengen würde und in diesem Bereich noch erheblicher Forschungsbedarf besteht.