# Bachelorarbeiten am Lehrstuhl 2

Prof. Dr. Amin Coja-Oghlan
amin.coja-oghlan@tu-dortmund.de

9. September 2021

*Effiziente Algorithmen bilden das Rückgrat aller IT-Systeme und gute Algorithmiker sind dementsprechend heiß begehrt. Am LS2 befassen wir uns mit der Entwicklung und Analyse von Algorithmen, mit Anwendungen in den verschiedensten Bereichen. Wenn Sie sich für eine Bachelorarbeit bei uns interessieren, können Sie wahlweise einen eher angewandten oder einen theoretischen Schwerpunkt setzen. Für ersteres sollten Sie Programmierkenntnisse und Experimentierfreude, für letzteres Begeisterung an einer analytischen Arbeitsweise mitbringen. Einige mögliche Themen in den Bereichen*

- *Kryptographie,*
- *Maschinelles Lernen,*
- *SAT-Problem,*
- *Epidemiologie,*
- *Komplexitätstheorie,*
- *Kombinatorik*

*sind im folgenden aufgelistet, wir sind aber immer auch für Vorschläge Ihrerseits offen.* ***Kontaktieren Sie mich gern per email. Ab dem 11. Oktober treffen Sie mich auch in der OH12, Raum 3.007 an.***

## Kryptographie

In der Kryptographie geht es darum, vertrauliche Informationen vor unbefugtem Zugriff zu schützen. Beispiele sind die verschlüss elte Kommunikation über öffentliche Netzwerke (z.B. “https”) sowie die Verschlüsselung von Daten auf Speichermedien. In der Kryptographie kommen vielfältige Algorithmen zur Ver- und Entschlüsselung zum Einsatz. Eine weitere Aufgabe besteht im Nachweis der Sicherheit von Verschlüsselungsverfahren, oder umgekehrt in der Untersuchung möglicher Angriffsstrategien.

### Thema: Das `EncFS`-Dateiverschlüsselungssystem

Moderne Betriebssysteme ermöglichen die Verschlüsselung ganzer Datenträger, um die Vertraulichkeit der Daten beispielsweise im Fall eines Diebstahls zu garantieren. Allerdings ist dieses Verfahren nicht immer anwendbar, etwa bei der Nutzung von Clouddiensten oder nicht selbst administrierten Speichermedien. Das `EncFS`-Verschlüsselungssystem erlaubt daher die nutzerseitige Verschlüsselung eines Verzeichnisses *innerhalb* eines existierenden Dateisystems. Ziel dieser

Bachelorarbeit ist die Untersuchung der kryptographischen Sicherheit sowie der algorithmischen Umsetzung solcher nutzerseitigen Verschlüsselungsverfahren. Welche Informationen über die gespeicherten Daten lecken nach außen (z.B. Dateigröße, Veränderungen), was sind mögliche Angriffspunkte und welche Alternativen zu existierenden Implementationen sind denkbar?

**Referenzen:**

- die `EncFS`-Projektseite [link]
- die `EncFS`-Seite in der ArchLinux-Wiki: [link]

### Thema: Blockchain-Technologie

Die Blockchain-Technik spielt nicht nur im Zusammenhang mit digitalen Währungen eine Rolle, sondern soll allgemein eine Art öffentlich zugreifbares aber dennoch sicheres "Logbuch" ermöglichen. In dieser Bachelorarbeit nehmen Sie die Algorithmen, auf denen die Blockchain-Technologie basiert, kritisch unter die Lupe und untersuchen auf dieser Grundlage mögliche Anwendungen dieser Technologie. Wie sicher ist die Blockchain-Technologie? Welcher algorithmische Aufwand ist für ihre Anwendung in verschiedenen Bereichen erforderlich?

**Referenzen:**

- Ein Artikel zu möglichen Anwendungen der Blockchain-Technologie im *Economist*. [link]
- S. Nakamoto: Bitcoin: a peer-to-peer electronic cash system. [link]

### Thema: Konstruktion von Einwegfunktionen

Einwegfunktionen sind Berechnungen, die einfach zu berechnen aber schwer zu invertieren sind. Beispielsweise ist es leicht, in einem (analogen) Telefonverzeichnis die Telefonnummer eines Teilnehmers herauszusuchen, aber schwer, den zu einer Telefonnummer gehörigen Teilnehmer zu ermitteln. Einwegfunktionen sind die Grundbausteine der Kryptographie. In dieser Bachelorarbeit geht es darum, neue Kandidaten für Einwegfunktionen zu entwickeln und auf die Probe zu stellen. Ausgangspunkt sind dabei schwere Berechnungsprobleme wie beispielsweise das Erfüllbarkeitsproblem.

**Referenzen:**

- Wikipedia-Artikel [link]
- O. Goldreich: Candidate One-Way Functions Based on Expander Graphs [link]

### Thema: Faktorisierungsalgorithmen

Bekannte Verschlüsselungsverfahren wie beispielsweise das häufig verwandte RSA-Verfahren beruhen auf der Annahme, daß es nur mit großem Aufwand möglich ist, eine große gegebene Zahl in ihre Primteiler zu zerlegen. Die Zahlen, von denen dabei die Rede ist, bestehen typischerweise aus einigen hundert oder einigen tausend Ziffern und haben keine "kleinen" Primteiler. Effiziente Faktorisierungsalgorithmen könnten also diese Verschlüsselungen brechen. In dieser Arbeit geht es um den Vergleich verschiedener Faktorisierungsalgorithmen. Welche Verfahren existieren, auf welchen mathematischen Techniken beruhen sie, wie effizient lassen sie sich implementieren und wie groß sind die Zahlen, die sich mit diesen Verfahren faktorisieren lassen? In der Konsequenz: wie sicher ist das RSA-Verfahren?

**Referenzen:**

- Wikipedia-Artikel [link]
- T. Kleinjung: Factorization of a 768-bit RSA modulus. [link]

### Thema: Quantencomputersichere Kryptographie

Während das Faktorisierungsproblem (eine gegebene Zahl in ihre Primteiler zu zerlegen) klassischen Computern schwer fällt, gibt es für dieses Problem auf Quantencomputern einen effizienten Algorithmus, den Shor-Algorithmus. Verschlüsselungen wie das RSA-Verfahren sind daher nicht sicher gegenüber Angriffen mit Quantencomputern. In dieser Arbeit befassen Sie sich mit alternativen Verschlüsselungsverfahren, die auch gegen Quantenangriffe sicher sind, sowie mit der Frage, welche weiteren "klassischen" Verfahren anfällig gegenüber solchen Angriffen sind.

**Referenzen:**

- Wikipedia-Artikel [link]
- What is Shor's factoring algorithm? [link]

## Maschinelles Lernen

Aufgrund spektakulärer Erfolge wie AlphaGo ist in den letzten Jahren das maschinelle Lernen wieder ins Zentrum der Informatikforschung gerückt. Algorithmen spielen dabei natürlich eine fundamentale Rolle, z.B. zum Trainieren eines neuronalen Netzwerkes. Ziel der folgenden Themen ist eine wissenschaftliche Auseinandersetzung mit maschinellem Lernen: was geht, was geht (noch) nicht? Was können verschiende Paradigmen leisten?

### Thema: Das Hopfield-Netzwerk

Das Hopfield-Netzwerk ist ein neuronales Netzwerk zum Abspeichern und Abrufen von Mustern. Derzeit ist die genaue Kapazität des Modells nicht bekannt und Ziel dieses Projektes ist es, Fortschritte zur Bestimmung der Kapazität zu machen. Heuristische Berechnungen deuten darauf hin, daß bei Überschreiten der Speicherkapazität die gesamte gespeicherte Information verlorengeht und das Netzwerk diese durch eine komplizierte Überlagerung zufälliger Muster überschreibt. Auch diese Hypothese soll untersucht werden.

**Referenzen:**

- Wikipedia-Artikel [link]
- M. Mezard: Mean-field message-passing equations in the Hopfield model and its generalizations. [link]

### Thema: Der Stochastic Gradient Descent-Algorithmus

Stochastic Gradient Descent ist der wichtigste Algorithmus im Bereich des "deep learning". Der relativ einfache Algorithmus wird zum Tranieren tiefer neuronaler Netze verwendet. Eine offene Frage ist, warum dieser Algorithmus trotz seiner Einfachheit in der Praxis so gut funktioniert. Eine weitere offene Frage ist, unter welchen Umständen das Verfahren scheitert. Einige Ansätze zum

Verständnis dieser Fragen wurden in der neuesten Literatur entwickelt. In dieser Arbeit sollen diese Verständnisansätze experimentell und/oder analytisch auf die Probe gestellt werden, um Forschritte beim Verständnis von Stochastic Gradient Descent zu erzielen.

**Referenzen:**

- Wikipedia-Artikel [link]
- S. Liu, D. Papailiopoulos, D. Achlioptas: Bad Global Minima Exist and SGD Can Reach Them. [link]

### Thema: Das Perceptron-Modell

Das Perceptron ist ein einfaches Modell eines neuronalen Netzwerkes, das jedoch noch nicht vollständig verstanden ist. Das Ziel ist eine binäre Klassifikation (z.B. “Hund” vs. “Katze”). In der neueren Forschung sind Fortschritte auf ein Verständnis dieses Modells gemacht worden, die Verbindungen zu Fragen in der statistischen Physik aufzeigen. Ziel dieser Arbeit ist ein besseres Verständnis des Modells, um seine genaue Leistungsfähigkeit besser einschätzen zu können.

**Referenzen:**

- Wikipedia-Artikel [link]
- J. Ding, N. Sun: Capacity lower bound for the Ising perceptron. [link]

### Thema: Clustering-Algorithmen

Ein grundlegendes Problem im maschinellen Lernen ist die Identifikation von Clustern in Daten. In dieser Bachelorarbeit sollen Clusterningalgorithmen für Netzwerkdaten untersucht und entwickelt werden. Dabei ist ein Netzwerk (in Form eines ungerichteten Graphen) gegeben und die Aufgabe besteht darin, Cluster ähnlicher Knoten in diesem Netzwerk zu identifizieren. Wichtig dabei ist, daß außer der Netzwerkstruktur keine Zusatzinformation gegeben ist; die Cluster sollen also allein auf Grundlage der Netzwerkinteraktionen identifiziert werden. Verschiedene algorithmische Ansätze sollen dabei anhand verschiedener Typen von Netzwerken untersucht werden.

**Referenzen:**

- E. Abbe: Community detection and stochastic block models. [link]
- C. Moore: The computer science and physics of community detection. [link]

### Thema: Hauptkomponentenanalyse

Bei der Hauptkomponentenanalyse (engl. principal component analysis) geht es darum, sehr große und umfangreiche Datensätze zu strukturieren und zu vereinfachen. Genauer gesagt, soll ein Großteil der vorhandenen Variablen durch Kombination von wenigen, sehr wichtigen Variablen (den Hauptkomponenten) beschrieben werden. Dabei soll möglichst wenig Information verloren gehen, indem mögliche Korrelationen innerhalb des Datensatzes genutzt werden. Eine Herausforderung bei der Hauptkomponentenanalyse ist es, daß es sehr viele Kombinationsmöglichkeiten der Variablen gibt, die im schlimmsten Fall alle durchprobiert werden müssen. Bei der “sparse” PCA, also einer “dünnen” Hauptkomponentenanalyse, wird aus diesem Grund gefordert, dass immer nur wenige Variablen genutzt werden dürfen. In diesem Projekt geht es darum, die Grundlagen

und Grenzen der dünnen Hauptkomponentenanalyse zu studieren und algorithmische Fragestellungen zu beantworten.

**Referenzen:**

- Wikipedia [link]
- Hui Zou, Trevor Hastie, Robert Tibshiran: Sparse principal component analysis. [link]

## Das SAT-Problem

Das Erfüllbarkeitsproblem ('SAT') ist eines der wichtigsten algorithmischen Probleme überhaupt, weil es als Unterproblem in einer Vielzahl von Anwendungen begegnet (z.B. Datenbankabfragen, Korrektheit von Programmen). Dabei geht es darum, für eine gegebene aussagenlogische Formel eine Belegung der Variablen zu finden, die die Formel erfüllt. Dieses Problem ist NP-vollständig, d.h. es gibt sehr wahrscheinlich keinen effizienten Algorithmus, der dieses Problem immer löst. Durch diese Erkenntnis verschwindet das SAT-Problem aber natürlich nicht von der Bildfläche. Stattdessen stellt sich die Aufgabe, dem SAT-Problem mit möglichst guten Heuristiken beizukommen. Darum geht es bei den folgenden Themenvorschlägen.

### Thema: SAT-lösen mit Conflict Driven Clause Learning

Die praktisch erfolgreichste Heuristik für das SAT-Problem ist derzeit Conflict Driven Clause Learning (CDCL). Ein Algorithmus versucht dabei zunächst, mit "trial and error" eine erfüllende Belegung zu konstruieren. Aus den dabei entstehenden Konflikten lernt der Algorithmus, und verwendet die so gewonnene Zusatzinformation bei seinem nächsten Versuch, die Formel zu lösen. In dieser Arbei soll ein einfacher CDCL-Algorithmus entwickelt und sein Verhalten experimentell und/oder analytisch untersucht werden. Auf welchen Typen von Instanzen ist CDCL erfolgreich? Welche Eigenschaften der zu lösenden Formel stellen den Algorithmus vor Schwierigkeiten?

**Referenzen:**

- J. Marques-Silva, I. Lynce, S. Malik: Conflict-Driven Clause Learning SAT Solvers. [link]
- Lingeling-SAT Software. [link]

### Thema: Lokale Suchalgorithmen für das SAT-Problem

Ein weiterer wichtiger algorithmischer Ansatz für das SAT-Problem ist lokale Suche. Der Algorithmus beginnt z.B. mit einer zufälligen Belegung der Variablen, die natürlich typischerweise keine Lösung der SAT-Formel ist. Anschließend "repariert" der Algorithmus nicht erfüllte Klauseln, möglicherweise wiederum unter Zuhilfenahme des Zufalls. Bekannte Beispiele lokaler Suchalgorithmen sind `Simulated Annealing` und `WalkSAT`. In diesem Projekt sollen derartige lokale Suchalgorithmen auf zufällig erzeugten Benchmark-Instanzen erprobt und untersucht werden.

**Referenzen:**

- Wikipedia. [link]
- Walksat Home Page. [link]

### Thema: Message Passing-Algorithmen

Eine neue Familie von Algorithmen für das SAT-Problem beruht auf "message passing". Die Idee ist, daß die Klauseln und Variablen der Formel einander Nachrichten senden, die immer wieder aufgrund der anderen Nachrichten, die die Variable/Klausel empfängt, aktualisiert werden. Wenn die Nachrichten konvergieren (d.h. sich nicht mehr wesentlich verändern), kann möglicherweise eine erfüllende Belegung abgelesen werden. Interessanterweise scheinen Message Passing-Algorithmen gerade mit solchen Instanzen gut zurechtzukommen, an denen andere Algorithmen scheitern. Ziel dieses Projektes ist es, Message Passing-Algorithmen auf verschiedenen Typen von SAT-Instanzen zu erproben und/oder zu analysieren.

**Referenzen:**

- L. Croc, A. Sabharwal, B. Selman: Survey Propagation revisited. [link]
- J. Pearl: Causality. [link]

## Algorithmen in der Epidemiologie

In der Covid-Pandemie boten computerbasierte (und damit algorithmenbasierte) Modelle der Ausbreitung des Virus eine wichtige Entscheiungsgrundlage für Eindämmungsmaßnahmen. Bei den folgenden Themen geht es um eine kritische Auseinandersetzung mit diesen Modellen. Welche Vorgänge können die verschiedenen Modelle abbilden? Erlauben sie eine nachvollziehbaere Erklärung der Vorhersagen? Wo gibt es Verbesserungsmöglichkeiten?

### Thema: Das OpenABM-Modell

Das an der Universität Oxford entwickelte OpenABM-Modell ist eines der wichtigsten Modelle in der Covid-Pandemie. Es zeichnet sich dadurch aus, daß es "open source", also frei zugänglich ist. Ziel dieses Projektes ist eine kritische Auseinandersetzung mit dem OpenABM-Modell. Das Modell ist agentenbasiert, d.h. die Vorhersagen beruhen auf der Simulation einer Vielzahl von Individuen, die sich anhand verschiedener Eigenschaften (Alter, Erwerbstätigkeit etc.) unterscheiden. Welche Annahmen gehen in die Modellierung ein, sind diese Annahmen notwendig und welchen Einfluss haben sie auf die Vorhersagen des Modells? Welche Effekte kann das Modell abbilden (z.B. psychologische Effekte), und wie vergleichen sich die Vorheragen im Verlauf dre Pandemie mit der Realität?

**Referenzen:**

- R. Hinch et al.: OpenABM-Covid19 - an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. [link]
- Covid19-Seite von Dirk Brockmann. [link]

### Thema: Ausbreitung von Epidemien auf Netzwerkmodellen

Im Gegensatz zu detaillierten agentenbasierten Modellen zielen Netzwerkmodelle darauf ab, Kontakte zwischen Individuen vereinfacht mit Hilfe einer Graphenstruktur zu modellieren. Der Vorteil ist, daß potentiell weniger Modellierungsparameter als bei einem agentenbasierten Modell erforderlich sind. Ein Nachteil ist, daß Kontakte zwischen Individuen mehr oder weniger als statisch angenommen werden. In diesem Projekt geht es um das Studium von Epidemien auf einfachen

Netzwerkmodellen. Ein sehr einfaches Modell ist etwa eine zweidimensionales Gitter, auf dem sich die Epidemie mehr oder weniger konzentrisch um ihren Ursprungsort ausbreitet. Was passiert, wenn man dem Modell zufällige Langstreckenverbindungen hinzufügt? Welche geometrischen Eigenschaften des Netzwerks bestimmen den Ausbreitungsprozess?

**Referenzen:**

- D. Brockmann, D. Helbing: The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. [link]
- B. Doerr, M. Fouz, T. Friedrich: Why rumors spread so quickly in social networks. [link]

### Thema: Netzwerkanalyse

Grundlage von netzwerkbasieren Epidemiemodellen ist natürlich die Kenntnis wichtiger Eigenschaften von Kontaktnetzwerken. In den letzten Jahren sind dazu zahlreiche Hypothesen entwickelt worden, bekannt unter Stichworten wie “six degrees of separation” oder “power laws”. In diesem Projekt sollen diese Hypothesen anhand öffentlich verfügbarer Netzwerkdaten kritisch hinterfragt werden. Erfüllen beispielsweise wissenschaftliche Kollaborationsnetzwerke die vorhergesagten Eigenschafen? Um dies zu untersuchen, müssen grundlegende Graphalgorithmen so implementiert werden, daß sie auch auf größere Datensätze anwendbar sind. Die Ergebnisse sollen dann mit statistischen Methoden ausgewertet werden.

**Referenzen:**

- Wikipedia. [link]
- R. Albert, A. Barabasi: Statistical mechanics of complex networks. [link]
- dblp computer science bibliography. [link]

## Komplexitätstheorie

Die Komplexitätstheorie zeigt die Grenzen effizienter Algorithmen auf. Die Kenntnis dieser Grenzen bewahrt uns einerseits davor, vergeblich “mit dem Kopf gegen die Wand zu laufen”. Andererseits, im positiven Sinne, ist die Komplexitätstheorie auch die Grundlage der Kryptographie, weil es in einer Welt, in der alles Probleme einfach sind, auch keine Geheimisse geben kann. Die folgenden Themen haben einen vorwiegend theoretische Einschlag.

### Thema: Das P≠NP-Problem

Die Komplexitätsklasse P bildet die Berechnungsprobleme ab, die sich mit effizienten Algorithmen lösen lassen. Die Klasse NP umfaßt zahlreiche Probleme von enormer praktischer Bedeutung, für die jedoch zum großen Teil keine effizienten Algorithmen bekannt sind. Ob für alle diese Probleme effiziente Algorithem existieren, ist vielleicht die wichtigste offene Frage der Informatik. Weithin wird eine negative Antwort, also P≠NP, erwartet. In diesem (rein theoretischen) Projekt soll der Forschungsstand zu dieser Frage zusammengestellt werden. Wie weit sind wir? Was sind die derzeit vielversprechendensten Zugänge zu diesem Problem? Welche Ansätze sind gescheitert (und woran)?

**Referenzen:**

- L. Hardesty: Explained: P vs. NP. [link]
- Clay Institute: P vs NP problem. [link]

### Thema: Komplexität und Proteinfaltung

Proteine sind lange Molekülketten, die als Grundbausteine der Biologie wichtige Funktionen in der Zelle wahrnehmen. Dabei falten sich Proteine in Anpassung an ihre Umgebung, um ihre Funktion erfüllen zu können. Korrekte und effiziente Proteinfaltung ist also überlebenswichtig. Der Faltungsprozess kann dabei durch ein relativ einfaches mathematisches Modell beschrieben werden. Eine komplexitätstheoretische Analyse zeigt jedoch bemerkenswerterweise, daß das Berechnungsprobleme, für ein gegebenes Protein die korrekte Faltung zu bestimmen, NP-schwer ist. Wie erklärt sich, daß in der Natur Proteine trotzdem zuverlässig schnell falten? Liegt es an der Co-Evolution der Organismen und der Proteine, oder an besonderen Eigenschafen von Proteine, die von Organismen wirklich gebildet werden?

**Referenzen:**

- Wikipedia-Artikel zur Proteinfaltung. [link]
- A. Fraenkel: Complexity of protein folding. [link]

## Kombinatorik

Kombinatorische Strukturen wie Graphen, Bäume, Polytope oder Codes bilden die mathematische Grundlage der Algorithmik. Die Entwicklung von Algorithmen beruht häufig auf kombinatorischen Erkenntnisse, und führt umgekehrt auf neue kombinatorische Fragestellungen. Die folgenden Themenvorschläge behandeln einige davon.

### Thema: Lokale Grenzwertsätze

Eine der wichtigsten Erkenntnisse in der Wahrscheinlichkeitstheorie ist, daß Summen unabhängiger Zufallsvariablen und sehr milden Annahmen gegen die Normalverteilung konvergieren. Diese Aussage ist bekannt als der “zentrale Grenzwertsatz”. In der Kombinatorik hat man es dabei häufig mit Zufallsvariablen zu tun, die irgendetwas zählen. Für solche Zufallsvariablen gilt manchmal eine noch genauere Aussage, nämlich ein “lokaler Grenzwertsatz”. In diesem Projekt soll es darum gehen, eine möglichst allgemeine Version des lokalen Grenzwertsatzes herzuleigen, die auf Anwendungen in der Kombinatorik zugeschnitten ist.

**Referenzen:**

- Wikipedia-Artikel. [link]
- Encyclopedia of Mathematics. [link]

### Thema: Algorithmische Regularität

Das Regularitätslemma ist ein zentrales Ergebnis der Kombinatorik. Es sagt aus, daß auch sehr große, komplizierte kombinatorische Strukturen immer als Überlagerung einer kleinen Zahl sehr regelmäßig strukturierter Strukturen beschrieben werden können. Aus algorithmischer Sicht stellt sich die Frage, wie eine solche reguläre Zerlegung möglichst effizient bestimmt werden kann. In

diesem Projekt geht es um die Entwicklung entsprechender Algorithmen für verschiedene kombinatorische Strukturen wie Graphen, Hypergraphen oder Wahrscheinlichkeitsverteilungen.

**Referenzen:**

- Abelpreis E. Szemeredi. [link]
- V. Rödl, M. Schacht: Regularity lemmas for graphs. [link]

### Thema: Rejection sampling

In diesem Projekt geht es um Algorithmen, die Objekte mit bestimmten erwünschten Eigenschafen zufällig erzeugen. Wir nehmen an, daß der Algorithmus einfache Münzwürfe durchführen kann, und das Ziel ist, daraus Algorithmen zur Erzeugung komplexerer Objekte zu entwickeln. In den letzten Jahren wurden neue Zugänge zu diesem Problem entwickelt, unter denen die Idee des "rejection sampling" hervorsticht. In diesem Projekt soll es darum gehen, dieses Verfahren auf neue kombinatorische Probleme zu übertragen und anzuwenden.

**Referenzen:**

- M. Jerrum: Fundamentals of partial rejection sampling. [link]
- B. Barak: Complexity of counting. [link]