Technische Universität Berlin
Fakultät IV: Elektrotechnik und Informatik
Institut für Softwaretechnik und Theoretische Informatik
Algorithmik und Komplexitätstheorie (AKT)

Community Detection-Algorithmen für Arzt-Patienten-Netzwerke

Bachelorarbeit

erstellt von Frank Ng

Matrikel: 316500

Berlin, 26.03.2018

zur Erlangung des Grades „Bachelor of Science“ (B. Sc.)
im Studiengang Informatik

Erstgutachter:	Prof. Dr. Rolf Niedermeier
Zweitgutachter:	Prof. Dr. Markus Brill
Betreuer:	Dr. André Nichterlein
	Prof. Dr. Rolf Niedermeier

Danksagung

Ich bin Prof. Dr. Rolf Niedermeier und vor allem Dr. André Nichterlein sehr dankbar für ihre Zeit, in der sie mich mit Rat unterstützt haben.

Für die Bereitstellung von Daten und Ressourcen danke ich Dr. Dominik von Stillfried und Thomas Czihal vom Zentralinstitut für die kassenärztliche Versorgung in Deutschland.

Außerdem danke ich Lotte Dammertz und Jessica Larkins für die abschließende Durchsicht der Arbeit.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Die selbständige und eigenhändige Anfertigung versichere ich an Eides statt.

Berlin, 26.03.2018 Frank Ng

Zusammenfassung

Komplexe Netzwerke weisen zumeist dichte Cluster auf. Die *Community Detection* ist eine Disziplin in der Sozialen Netzwerkanalyse, um solche Cluster, auch *Community* genannt, zu identifizieren [For10]. Es existieren zahlreiche Community Detection-Methoden. Im Bereich des Gesundheitswesens sind Community Detection-Verfahren interessant, um dichte Arztnetze zu identifizieren.

In dieser Arbeit wird der Algorithmus *Arztnetz-Detection* vorgestellt, um Arztnetze mit festen Vorgaben zu identifizieren. Der Algorithmus besteht aus vier Phasen und verwendet dafür unter anderem die Community Detection-Algorithmen *LPAwb+* von Beckett [Bec16] und *Louvain* von Blondel u.a. [Blo+08] für die ersten zwei Phasen. In der dritten Phase werden die zu großen Communities und in der vierten Phase die zu kleinen Communities derart bearbeitet, sodass sie den Vorgaben entsprechen.

Das Hauptaugenmerk des Algorithmus liegt darin, mit Hilfe von Community Detection-Verfahren dichte Cluster zu finden und mit Hilfe anderer im Algorithmus eingebauten Mechanismen die festen Vorgaben an Größe der Cluster und die Unterschiedlichkeit der darin enthaltenen Knoten für die Arztnetz-Identifizierung zu erfüllen.

Der hier vorgestellte Algorithmus schafft es für 203 Testnetzwerke mit durchschnittlich 154 Ärzte und 5130 Patienten im Schnitt 150 Ärzten einem Arztnetz zuzuordnen, sodass die Arztnetze auch alle Vorgaben im Rahmen der Vorgaben der Kassenärztlichen Bundesvereinigung zu erfüllen. Die identifizierten Arztnetze weisen zudem eine höhere Dichte und einen höheren Cluster-Koeffizienten als die im gesamten Netzwerk aus. Die Laufzeit beträgt im Schnitt 36 Minuten, was den Algorithmus für Testnetzwerke in dieser Größenordnung praxistauglich macht.

Es zeigt sich jedoch auch, dass die bipartite Community Detection mit der Optimierung von Barbers Modularität bei der Arztnetz-Findung nicht geeignet ist und hierfür Bearbeitungsbedarf besteht.

Abstract

Complex networks often exhibit dense clusters. Community detection is a discipline in the field of social network analysis which serves to identify these clusters or, as they are sometimes also referred to, communities [For10]. Moreover, in the healthcare sector community detection methods can be used to identify physician communities.

This thesis introduces the algorithm *Arztnetz-Detection* which identifies physician communities with fixed requirements.

The algorithm consists of four phases and for the first and second phases it uses Beckett's *LPAwb+* [Bec16] and Blondel et al's *Louvain* [Blo+08] as community detection-algorithms. In the third phase too big communities and in the fourth phase too small communities will be adapted so that they satisfy the requirements.

The main focus of the algorithm is to identify dense clusters via community detection-methods and by using the other mechanisms built into the algorithm satisfy the fixed requirements pertaining to tze size of the clusters and the diversity of the knots contained in them for the physician network detection.

The introduced algorithm is able to assign from 203 test networks with on average 154 physicians and 5130 patients on average 150 physicians to a physician community so that every physician community satisfies the requirements introduced by the Kassenärztliche Bundesvereinigung (KBV).

The identified physician communities show a higher density and cluster-coefficient than the desity and cluster-coefficient of the whole networks. The computing time is on average 36 minutes which makes the algorithm useful in practice, at least for networks of similar sizes as mentioned above.

Furthermore, it seems that optimizing Barber's modularity for the bipartite community detectin is not appropriate for identifying physician communities and thus needs some modification.

Inhaltsverzeichnis

1 Einführung

Kooperationen sind wichtig, da sie komplexe Aufgaben durch Wissens- und Ressourcenaustausch sowie Arbeitsaufteilung effizienter machen können. Vor allem im Bereich Gesundheitswesen sind Kooperationen von Ärzten erwünscht.

Auf Grund des technischen und wissenschaftlichen Fortschrittes spezialisieren sich Ärzte immer mehr. Patienten werden in den häufigsten Fällen für ein Krankheitsbild von mehreren Ärzten verschiedener Fachgruppen behandelt. Zur Verbesserung der Patientenversorgung gehört somit die Kooperation zwischen Ärzten. Die Idee von Arztnetzen ist nicht neu, es existieren schon einige Arztnetze in Deutschland. Es besteht jedoch Potential, weitere Kooperationen von Ärzten zu fördern.

In dieser Arbeit wird der Algorithmus *Arztnetz-Detection* entwickelt und getestet, der potentielle Arztnetze aus den vertragsärztlichen Abrechnungsdaten herausgeben kann. Dazu betrachtet er die Arzt-Patienten-Kontakte als Verbindungen und baut daraus ein bipartites Netzwerk. Mit schon bestehenden Methoden, Community Detection, aus der sozialen Netzwerkanalyse wird der Algorithmus in vier Phasen Communities verschiedener Qualitätsklassen identifizieren, die die Gütemaße für Arztnetze laut den Rahmenvorgaben der Kassenärztlichen Bundesvereinigung erfüllen.

Neben der Erfüllung der Gütemaße ist die Laufzeit sowie die vollständige Zuordnung der Ärzte in Arztnetze wichtig. Ein nicht notwendiges Kriterium ist die Dichte der Arztnetze, was jedoch gewollt ist, da dies neben den harten Kriterien sich von einer zufälligen Zusammensetzung von Ärzten unterscheidet. Daher werden Community Detection - Verfahren verwendet.

Der hier entwickelte Algorithmus zeigt Potentiale, dichte Arztnetze mit den vordefinierten Gütemaßen zu identifizieren. Für die Hälfte der vorhandenen Kreise konnte in einer praxistauglichen Zeit die Zuordnung vorgenommen werden. Mögliche Verbesserungen werden am Ende dieser Arbeit thematisiert.

1.1 Aufgabenstellung

Aus einem bipartiten Arzt-Patienten-Netzwerk sollen Arztnetze identifiziert werden, die die harten Kriterien erfüllen. Die harten Kriterien sind die Mindest- und Höchstanzahl an Ärzten in einem Arztnetz sowie die Mindestanzahl an verschiedenen Fachgruppen der Ärzte. Diese harten Kriterien werden hier als explizite Gütemaße bezeichnet, weil sie explizit angegeben werden. Diese können also variieren und werden im Kapitel 3 detaillierter beschrieben.

Neben den harten Kriterien gibt es die weichen Kriterien wie die Dichte eines Arztnetzes. Die Beziehungen zwischen den Ärzten eines Arztnetzes sollen möglichst eine höhere Dichte aufweisen als die Beziehungen aller Ärzte im Gesamtnetzwerk. Durch die verwendeten Community Detection-Methoden soll dies erreicht werden.

Die Aufgabe besteht im Zerlegen des bipartiten Netzwerkes in möglichst viele geeignete Cluster, die die harten Kriterien erfüllen und somit die Arztnetze repräsentieren. Die Aufgabe des Algorithmus kann folgendermaßen definiert werden:

Arzt-Patienten - Community Detection
Eingabe: Bipartites Arzt-Patienten-Netzwerk AP und Gütemaße B.

- Das Arzt-Patient-Netzwerk $AP = (A \cup P, E, FG, W)$ ist ein Tupel aus der Vereinigung von der Arztmenge und der Patientenmenge $A \cup P$, der Kantenmenge E, der Fachgruppenmenge FG und dem Kantengewicht
- $E = \{(a, p), a \in A, p \in P\}$ ist die Menge an Kanten. Es besteht lediglich eine Verbindung von Arzt $a \in A$ zum Patienten $p \in P$.
- FG enthält jede Fachgruppe als eine eindeutige Zahl. Jeder Arzt besitzt eine eindeutige Fachgruppe und die Zuordnung kann folgendermaßen definiert werden: $fg : A \to FG$.
- W enthält die Gewichte jeder Kante. Das Kantengewicht ist die Anzahl an Behandlungsfällen des Patienten beim Arzt. Die Zuordnung wird folgendermaßen definiert: $w : E \to W$, wobei $w \in \mathbb{N}$.
- $B = (b_1, b_2, b_3)$ ist eine Auflistung von drei Eigenschaften, die jedes Cluster erfüllen soll, um als für Arztnetze geeignet zu sein. b_1 ist die Mindestanzahl an Ärzten. b_2 ist die Höchstanzahl an Ärzten. b_3 ist die Mindestanzahl an distinkten Fachgruppen.

Aufgabe: Finde ein geeignetes Clustering aus Ärzten für AP, sodass alle Cluster die Gütemaße aus B erfüllen. Nicht alle Ärzte müssen einem Cluster zugeordnet werden.

1.2 Literaturüberblick

Zwar wurde bisher kein Algorithmus entwickelt, um potentielle Arztnetze zu erstellen, jedoch sind einzelne Themenbereiche nicht ganz neu und es existieren sowohl Quellen als auch Arbeiten dazu.

Arztnetze: Es existieren schon Arztnetze in Deutschland und diese haben in der Agentur Deutscher Arztnetze einen Interessenvertreter[1]. Die Agentur Deutscher Arztnetze beschreibt zwei Formen von Arztnetzen. Zum Einem ein Arztnetz als Modellvorhaben (§ 63 ff SGB V) und zum Anderen das Arztnetz mit Strukturverträgen (§ 73 a SGB V).

Beim Modellvorhaben können Ärzte mit den Krankenkassen Direktverträge abschließen. Kassenärztliche Vereinigungen können mit den Krankenkassen die Rahmenbedingungen für das Modellvorhaben vereinbaren. Das Ziel ist die Optimierung von Verfahren, Vergütung, Finanzierung und Organisation. Strukturverträge können mit den Kassenvereinigungen abgeschlossen werden. Ziel ist hierbei die Optimierung der Versorgungs- und Vergütungsstrukturen im ambulanten Bereich. In den Strukturverträgen können auch eigene Vergütungsstrukturen vereinbart werden.

Auch können Kassenärztliche Vereinigungen laut §87 b SGB V Praxisnetze fördern[2]. Die Rahmenvorgabe der Kassenärztlichen Bundesvereinigung für die Anerkennung von Praxisnetzen nach § 87b Abs. 4 SGB V existiert seit 2013[3].

Netzwerke: Bei der Analyse von komplexen Netzwerken spielt das Auffinden von dichten Clustern, die zueinander relativ wenige Verbindungen haben, eine große Rolle. Diese dichten Cluster werden auch Community genannt und können wiederum als einzelne Teilnetzwerke analysiert werden. Zum Beispiel hat ein Freundschaftsnetzwerk wie Facebook eine deutliche Community-Struktur [Uga+11].

Newman [New10] bietet eine gute Einführung in die Netzwerktheorie. In seinem Buch finden sich Erklärungen über verschiedene Netzwerke. Dazu gehören technologische Netzwerke wie ein Telefonnetzwerk, soziale Netzwerke, Informationsnetzwerke wie das Internet und biologische Netzwerke wie zum Beispiel Proteinnetzwerke.

Newman führt verschiedene Eigenschaften von Netzwerken und seinen Akteuren sowie deren mathematische Berechnungen auf. Ein wichtiger Teil für diese Arbeit ist die Beschreibung von Graph-Partitionierung und ihrem Teilgebiet: Community Detection.

Netzwerke können verschiedene Typen von Elementen haben. In diesem Fall existieren zwei Typen: Arzt und Patient.

Community Detection: Girvan und Newman [GN02] beschreiben die Community-Struktur von sozialen und biologischen Netzwerken und entwickeln eine Methode, um diese Communities zu entdecken. Newman entwickelt Community Detection - Methoden auf Basis der Modularität, die auch in vielen anderen Algorithmen verwendet werden.

[1] http://deutsche-aerztenetze.de/
[2] http://www.kbv.de/media/sp/PraxisWissen_Praxisnetze_web.pdf
[3] http://www.kbv.de/media/sp/Rahmenvorgabe_Anerkennung_Praxisnetze_Ausfertigung.pdf

Fortunato hat mehrere Arbeiten zum Thema Community Detection veröffentlicht. Eine Arbeit von ihm [For10] listet ausführliche Methoden zur Erkennung von Communities in Graphen auf. 2016 haben Fortunato u.a. [FH16] eine weitere Arbeit über Community Detection veröffentlicht, in der auch viele Methoden von anderen Autoren beschrieben wurden.

Barber entwickelt die Modularität für bipartite Netzwerke und beschreibt das Verfahren von Community Detection in bipartiten Netzwerken [Bar07].

In einem Projektbericht untersuchen Bodoia u.a. [BM14] verschiedene Paradigmen zur Identifizierung von Communities in bipartiten Netzwerken.

Community Detection im Gesundheitswesen: Es existieren eher im internationalen als im deutschen Raum Arbeiten über Community Detection im Gesundheitswesen.

Landon u.a. [Lan+13] haben in 51 Regionen in den USA untersucht, wie sich ein Community-basiertes Netzwerk, das durch den Community Detection-Algorithmus von Newman [New04] erstellt wurde, und ein Krankenhaus-basiertes Netzwerk, das nach formalen Kriterien erstellt wurde, sich voneinander unterscheiden.

Barnett u.a. [Bar+12] erstellen Ärzte-Netzwerke in verschiedenen Krankenhäusern ebenso durch gemeinsame Patienten. Die Auswirkungen verschiedener Netzwerkstrukturen werden im Hinblick auf die Krankenhauskosten untersucht.

In Deutschland existiert eine Arbeit von von Stillfried [Sti+17], in der Ansätze zum Aufbau von virtuellen Ärztenetzwerken beschrieben werden und sich von geologisch abgetrennten Netzwerken unterscheiden.

1.3 Struktur der Arbeit

Im Kapitel 2 **Grundlagen** werden die Grundzüge des deutschen Gesundheitswesens erklärt. Außerdem werden die Routinedaten aus den kassenärztlichen Abrechnungsdaten, aus denen die Arzt-Patienten-Netzwerken gebildet werden, umrissen. Des weiteren werden bipartite Netzwerke sowie die wesentlichen Punkte bei der Community Detection erläutert.

Im Kapitel 3 **Gütemaße** wird beschrieben, welche weichen und harten Kriterien verwendet werden, um geeignete Communities für Arztnetze zu definieren. Die weichen Kriterien sind die Modularitäten, die für die Community Detection-Algorithmen verwendet werden. Die harten Kriterien sind für die Definition von Arztnetzen von Bedeutung.

Im Kapitel 4 **Community Detection - Algorithmen** wird zuerst der eingesetzte *Louvain*-Algorithmus und *LPAwb+*-Algorithmus zur Auffindung von Communities in den unipartiten und bipartiten Netzwerken beschrieben. Im nächsten Abschnitt wird der eigene Algorithmus *Arztnetz-Detection* detailliert beschrieben, der im Rahmen dieser Arbeit entwickelt wird.

Im Kapitel 5 **Arzt-Patienten - Community Detection** wird der Algorithmus *Arztnetz-Detection* getestet. Als Testdaten werden reale Arzt-Patienten-Netzwerke aus verschiedenen Ortskreisen erstellt. Um diese zu erstellen werden die vertragsärztlichen Abrechnungsdaten verwendet. Die Ergebnisse werden präsentiert. Außerdem werden mögliche Optimierungen des Algorithmus diskutiert.

Im Kapitel 6 **Fazit** wird das Ergebnis zusammengefasst. Das Hauptaugenmerk richtet sich darauf, inwieweit der Algorithmus den gestellten Ansprüchen genügt und welche Punkte verbessert werden müssen.

2 Grundlagen

Im Abschnitt 2.1 wird das Gesundheitswesen in Deutschland kurz erläutert. Im Abschnitt 2.2 wird beschrieben, welche Informationen administrative Daten bieten und wie daraus Arzt-Patienten-Netzwerke gewonnen werden. Abschnitt 2.3 beschreibt bipartite Netzwerke. Im Abschnitt 2.4 wird die Community Detection beschrieben.

2.1 Gesundheitswesen in Deutschland

Das deutsche Gesundheitssystem ist in die ambulante und die stationäre Versorgung aufgeteilt. Während im stationären Bereich die Krankenhäuser die Versorgung der Patienten übernehmen, werden im ambulanten Bereich die Leistungen außerhalb der Krankenhäuser von niedergelassenen Ärzten, Psychotherapeuten und Zahnärzten erbracht. [BB14]

Aus den Berechnungen des Zentralinstituts für die kassenärztliche Versorgung in Deutschland (Zi) auf Basis der bundesweiten Abrechnungsdaten nach §295 im 5. Sozialgesetzbuch (SGB) und der fallpauschalenbezogenen Krankenhausstatistik (DRG-Statistik) des Statistischen Bundesamts ergeben sich für das Datenjahr 2016 folgende Kennzahlen:

- Rund 71 Millionen GKV-Patienten besuchten mindestens einen ambulanten Arzt.
- Es gab rund 600 Millionen ambulante Behandlungsfälle, wobei ein dokumentierter Behandlungsfall mehrere Besuche eines Patienten in einem Quartal erfassen kann.
- Es behandelten rund 181.000 niedergelassene Ärzte.
- Es gab rund 126.000 Praxen, davon waren rund 90.000 Einzelpraxen.
- Die Kosten für die ärztlichen Leistungen betrugen rund 40 Mrd. Euro.

Viele niedergelassene Ärzte sind selbstständig, haben ihre eigene Praxis oder teilen sich eine Praxis mit Kollegen. Jeder niedergelassene Arzt ist automatisch Mitglied der kassenärztlichen Vereinigung (KV) in seiner Region. Die jeweilige KV repräsentiert ihre Mitglieder und verhandelt mit den gesetzlichen Krankenkassen die Vergütung der einzelnen Leistungen. Die Krankenkassen zahlen in den Gesundheitsfond ein, aus welchem die Vergütung der Ärzte bestimmt wird.

Zu jedem Quartal sendet der Arzt seiner jeweiligen KV seine administrativen Daten, die alle nötigen Informationen wie die von Patienten, Diagnose, Leistungen und vom behandelnden Arzt enthalten. Diese administrativen Daten dienen dazu, für jeden Arzt seinen Anteil aus dem Gesundheitsfond zu ermitteln.

Diese administrativen Daten werden für die weitere Forschung pseudonymisiert gespeichert. Datenhalter sind neben den 17 Kassenärztlichen Vereinigungen (16 Bundesländer haben jeweils eine KV, Nordrhein-Westfalen hat zwei: KV Nordrhein und KV Westfalen-Lippe) und die Kassenärztliche Bundesvereinigung (KBV) auch das Zentralinstitut für die kassenärztliche Versorgung in Deutschland (Zi). [SEC14]

2.2 Arzt-Patienten-Netzwerke

Als Datengrundlage liegen pseudonymisierte vertragsärztliche Abrechnungsdaten vor. [Got+14] In diesen können Ärzte, Patienten und Behandlungsfälle unterschieden werden. Unter Anderem sind Daten wie der Wohnort des Patienten, die Fachgruppe des Arztes und die Feststellung, ob ein Patient von einem anderen Arzt überwiesen wurde, enthalten.

In jedem Behandlungsfall wird der behandelnde Arzt und der Patient erfasst, womit eine Verbindung zwischen Ärzten und Patienten identifiziert werden kann. Folglich kann ein Arzt-Patienten-Netzwerk gebildet werden, in welchem die Ärzte und die Patienten die zwei verschiedenen Knotentypen und der Behandlungsfall die Kante zwischen den Knoten darstellen. Die Anzahl der Behandlungsfälle zwischen Arzt und Patient kann als das Gewicht einer Kante aufgefasst werden.

Es werden mehrere Arzt-Patienten-Netzwerke nach Wohnkreisen erstellt. Dazu dienen der Wohnort des Patienten zum Zeitpunkt der Behandlung. Beim Arzt ist lediglich seine KV enthalten, bei der seine Leistungen abgerechnet werden. Um dem Arzt den Wohnkreis zuzuordnen, werden die Wohnorte seiner Patienten ermittelt. Dem Arzt wird der Wohnort zugeordnet, in dem die Mehrheit seiner Patienten lebt.

Weitere Metadaten wie Leistungen und Diagnosen sowie Alter und Geschlecht des Patienten werden in dieser Arbeit nicht verwendet, können aber Aufschluss über die Bildung von bestimmten Arzt-Patienten-Netzwerken geben.

2.3 Bipartite Netzwerke

Netzwerke, die aus zwei unterschiedlichen Typen von Netzwerkelementen bestehen, sind bipartite Netzwerke. Beispiele sind Schauspieler und Filme, Forscher und Publikationen, Käufer und Waren, oder in diesem Fall: Ärzte und Patienten. In bipartiten Netzwerken ist die Besonderheit, dass Knoten desselben Typs zueinander keine Kanten haben.

Für die Identifizierung von Gemeinschaften werden bipartite Netzwerke [BE97][Bor09] oft in einfachere unipartite Netzwerke überführt. Dies geschieht durch die Verbindung von Akteuren, die eine Gemeinsamkeit haben. Zum Beispiel werden Schauspieler miteinander verbunden, wenn sie in den selben Filmen mitgespielt haben. Informationen über die Filme gehen durch diesen Prozess verloren. Guimerà u.a. [GSPA07] zeigen, dass Analysen von ungewichteten unipartiten Projektionen zu unzuverlässigen oder falschen Resultaten führen können.

Aus den hier verwendeten administrativen Daten lassen sich bipartite Netzwerke aus Ärzten und Patienten, die durch den Besuch eines Patienten beim Arzt verbunden werden, erstellen.

2.4 Community Detection

Reale Netzwerke haben zumeist eine klare Community-Struktur. Als Beispiel kann das Facebook-Netzwerk genommen werden, das ein dünner Graph ist mit viele dichten Teilgraphen beziehungsweise Communities enthält. In diesen Communities existieren zwi-

schen den Akteuren viele Verbindungen, während relativ wenige Verbindungen zwischen Akteuren unterschiedlicher Communities bestehen [Uga+11].

Weil Arztnetze möglichst aus Ärzten bestehen sollten, die laut den Abrechnungsdaten viele gemeinsame Patienten haben, ist hier die Community Detection sehr interessant.

Newman [New10] beschreibt Community Detection als eine Methode, um natürliche Abgrenzungen in einem komplexen Netzwerk zu finden. Denn im Gegensatz zum klassischen Problem der Graphpartitionierung, werden bei der Community Detection die Größe und Anzahl an Clustern nicht vorab definiert. Die Communities können je nach Netzwerk und Methode in Anzahl und Größe variieren [GN02].

Beide Probleme unterscheiden sich zudem in ihren Zielen. Primär wird durch die Graphpartionierung ein Netzwerk in kleinere, besser kontrollierbare Teile getrennt. Community Detection erschafft klare Trennungen. Das eigentliche Ziel der Community Detection ist jedoch, ein besseres Verständnis über die natürliche Struktur eines Netzwerkes zu entwickeln.

Community Detection wird meist auf unipartiten Netzwerken durchgeführt. Auch bipartite Netzwerke werden in unipartite überführt, um darauf eine Community Detection auszuführen. Guimerà [GSPA07] zeigt, dass dieser Schritt zu falschen Ergebnissen führen kann.

Bodoia u.a. [BM14] beschreiben drei Paragdimen, um Community-Strukturen in bipartiten Netzwerken zu identifizieren:

- Paradigma 1: typische Community Detection-Algorithmen, die nicht explizit für bipartite Netzwerken entwickelt wurden
- Paradigma 2: für bipartite Netzwerke entwickelte Community Detection-Algorithmen
- Paradigma 3: Projektion der bipartiten Netzwerke in unipartite, gewichtete Netzwerke, um diese mit den Algorithmen aus Paradigma 1 zu bearbeiten

Letztlich ist die Community Detection für bipartite Netzwerke noch nicht so gut erforscht, wie für unipartite Netzwerke. Trotzdem kommt das Thema bipartites Netzwerk und seine Partitionierung immer mehr ins Blickfeld von Forschern verschiedener Disziplinen. Jedes Jahr finden sind neue Algorithmen oder Ansätze, um die Community-Struktur in einem bipartiten Netzwerk zu identifizieren [Bec16; LCJ14; LM10; PCK16; ZA15]

3 Gütemaße

In diesem Kapitel werden zwei Arten von Gütemaßen beschrieben. Zum einen gibt es die **impliziten** Gütemaße und zum anderen gibt es die **expliziten** Gütemaße. Die folgenden Modularitäten gelten somit als implizite Gütemaße, die nicht vorher festgelegt werden, jedoch für die Erstidentifizierung von dichten Clustern wichtig sind. Das erste Ziel ist die Identifizierung der Community-Struktur im Netzwerk. Eine Partitionierung eines Netzwerkes in dichte Cluster kann als Community-Struktur gesehen werden.

In dieser Arbeit werden die *Louvain Methode* [Blo+08] und der *LPAwb+* Algorithmus von Beckett [Bec16] verwendet, um die Community-Struktur zu identifizieren.

Die *Louvain Methode* identifiziert die Community-Struktur bei unipartiten Netzwerken, indem sie versucht, die Modularität zu optimieren. Die Modularität wird im Unterkapitel 3.2 beschrieben.

LPAwb+ ist eine Methode, um eine Community-Struktur in bipartiten Netzwerken zu finden finden. Sie erreicht ihr Ziel durch die Optimierung der Modularität von Barber [Bar07], eine bipartite Version der Modularität. Barbers Modularität wird im Unterkapitel 3.3 weiter erläutert.

Nachdem die dichten Cluster gefunden wurden, werden sie auf die expliziten Gütemaße geprüft. Die expliziten Gütemaße sind wichtig, um Cluster herauszufiltern, die als Arztnetze geeignet sind. Diese Gütemaße sind die minimale und maximale Anzahl an Ärzten in einer Community sowie die minimale Anzahl an verschiedenen Fachgruppen der Ärzte (siehe Abschnitt 3.1).

Der Unterschied zwischen den impliziten und den expliziten Gütemaßen ist, dass die impliziten Gütemaße wie Modularität oder Barbers Modularität nicht angegeben werden, aber implizit in den Community Detection - Algorithmen verwendet werden, um die bestmögliche Community-Struktur zu erhalten.

Die expliziten Gütemaße werden als Parameter in den Algorithmus eingegeben. Der Benutzer des Algorithmus beeinflusst mit seinen Wunschangaben den Output.

3.1 Explizite Gütemaße: Größe des Arztnetzes und Anzahl verschiedener Fachgruppen

Laut der Rahmenvorgabe der Kassenärztlichen Bundesvereinigung für die Anerkennung von Praxisnetzen nach § 87b Abs. 4 SGB V [4] gibt es Voraussetzungen, damit Arztnetze besonders förderungswürdig sind.

Eine Voraussetzung ist die Teilnahme von mindestens 20 und höchstens 100 vertragsärztlichen Praxen. Für diese Arbeit werden statt Praxen Ärzte betrachtet. Für den Algorithmus zur Auffindung von Arztnetzen spielt dies keine Rolle. Das erste explizite Gütemaß ist also die **minimale Anzahl an Ärzten** und das zweite explizite Gütemaß ist die **maximale Anzahl an Ärzten** in einem Arztnetz. Weiter wird geschrieben, dass mindestens 3 Fachgruppen in einem Arztnetz vertreten sein müssen. Somit ist das dritte explizite Gütemaß die **minimale Anzahl an verschiedenen Fachgruppen**.

[4] http://www.kbv.de/media/sp/Rahmenvorgabe_Anerkennung_Praxisnetze_Ausfertigung.pdf

Alle anderen Voraussetzungen sind nicht oder nur sehr schwer aus den Abrechnungsdaten entnehmbar. Daher werden diese nicht als Gütemaße betrachtet.

Diese expliziten Gütemaße sind interessant, weil sie aus einer realen Rahmenvorgabe resultieren. Dadurch lässt sich die Anwendbarkeit dieser Methode auf reale Strukturen untersuchen. Ein Arztnetz kann nur gebildet werden, wenn die realen Voraussetzungen erfüllt werden. Eine perfekte Clusterung eines Netzwerkes in Communities und eine optimale Modularität reichen als Kriterien nicht aus.

Somit sind diese expliziten Gütemaße die Bedingungen, die erfüllt werden müssen, wobei die impliziten Gütemaße aus Abschnitten 3.2 und 3.3 möglichst hoch beziehungsweise zutreffend sein sollen, aber keine Ausschluss- oder Einschlusskriterien sind.

Es bedeutet jedoch nicht, dass nicht in Zukunft weitere statistische Gütemaße mit einfließen sollen. Die Auswertung der Arztnetze im Kapitel 5 soll auch dazu dienen, eventuell weitere Gütemaße zu definieren. Dafür ist jedoch die Analyse des Ist-Zustandes wichtig.

3.2 Implizites Gütemaß I: Modularität

Die Modularität wurde von Newman und Girvan [GN02] eingeführt. Durch die Optimierung der Modularität kann das Netzwerk so aufgeteilt werden, sodass die Cluster möglichst dicht sind. Außerdem dient sie in vielen Arbeiten als Metrik, um verschiedene Community Detection - Ansätze zu vergleichen.

Die Modularität wird folgendermaßen definiert:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(g_i, g_j) \tag{1}$$

$$P_{ij} = \frac{k_i k_j}{2m} \tag{2}$$

Dabei ist m die Anzahl aller Kanten und A_{ij} die Adjazenzmatrix, die das Netzwerk repräsentiert. Jedes Element a_{ij} in A_{ij} zeigt für das Knotenpaar (i, j), ob es eine Verbindung zwischen diesen besteht. Falls es keine Verbindung zwischen i und j besteht, dann ist $a_{i,j} = 0$, sonst 1 oder w, falls die Kanten auch gewichtet werden.

P_{ij} ist das Nullmodell, das einen durchschnittlichen Graphen repräsentiert, der einige Eigenschaften des Originalgraphen behält, wie zum Beispiel die Anzahl an Knoten und der Grad der Knoten. Die Idee ist, dass durch die Randomisierung des Graphen die Community-Struktur zerstört ist. Je mehr sich nun der Originalgraph von dem Nullmodell unterscheidet, desto modularer ist der Originalgraph. Eine hohe Modularität weist auf eine gute Community-Struktur hin.

So zeigt hier das Element p_{ij} von P_{ij} nicht an, ob es eine Verbindung zwischen jedem Knotenpaar i, j gibt, sondern die Wahrscheinlichkeit einer Verbindung zwischen Knoten i und Knoten j. Sei k_i der Grad vom Knoten i und k_j der Grad vom Knoten j, dann wird die Wahrscheinlichkeit und somit jedes Element von $P_{i,j}$ wie in der Gleichung 2 ausgerechnet.

Der Term $(A_{ij} - P_{ij})$ in der Gleichung (1) wird negativ, wenn i und j mit anderen Knoten verbunden sind, aber untereinander nicht. Besteht eine Kante zwischen i und

j, dann wird der Term umso größer, je weniger Kanten beide Knoten insgesamt haben und somit die Wahrscheinlichkeit gering wäre, dass sich diese beiden Knoten in einem zufälligen Graphen verbinden.

Die Kronecker-Delta-Funktion $\delta(g_i, g_j)$ ist gleich 1, wenn beide zu betrachtenden Knoten in einer Community sind, ansonsten ist $\delta(g_i, g_j) = 0$. Somit stellt sie sicher, dass Verbindungen von Knoten innerhalb einer Community und die Verbindungen von diesen Knoten nach außen betrachtet werden.

Zusammengefasst ist die Modularität Q die Summe aller Differenzen zwischen den tatsächlichen Verbindungen von Knoten innerhalb einer gegeben Community und den zu erwarteten Verbindungen in einem Nullmodell. Je mehr Verbindungen innerhalb von Communities vorhanden sind, desto höher ist Q. Wenn es weniger Kanten als erwartet gibt, dann ist Q negativ. Wenn es nur eine oder keine Community gibt, dann ist $Q = 0$.

Die Modularität hat jedoch auch einige Schwächen. Fortunato [FB07] beschreibt die *Resolution Limitation* in der Modularität, die dafür sorgt, dass sehr kleine Gemeinschaften, sogar Cliquen, zu einer größeren Gemeinschaft verbunden werden, auch wenn diese zueinander wenige Verbindungen haben. Guimerà [GSPA07] entdeckt Partitionen mit hohen Modularitätswerten bei Zufallsgraphen, was der Idee der Modularität widerspricht.

Aufgrund ihrer Einfachheit basieren trotz dieser Limitationen viele Algorithmen wie zum Beispiel die hier genutzte Louvain-Methode [Blo+08] auf der Modularität. Es können aber auch andere Verfahren eingesetzt und getestet werden.

3.3 Implizites Gütemaß II: Barbers Modularität

Bipartite Netzwerke haben im Gegensatz zu unipartiten Netzwerken die Bedingung, dass Knoten desselben Typs keine Kante zueinander haben. Diese Bedingung wurde in der Modularität von Barber berücksichtigt.

Die Gleichung (1) für unipartite Netzwerke würde im Term $(A_{ij} - P_{ij})$ das Nichtvorhandensein von Kanten zwischen Knoten innerhalb derselben Community mit einem negativen Wert bestrafen. Nun sind jedoch per Definition eines bipartiten Graphen die Knoten des selben Typs nicht miteinander verbunden. Folglich würde eine Community mit vielen Knoten die Modularität negativ beeinflussen.

Barber [Bar07] definiert eine Modularität für bipartite Netzwerke, um dieses Problem zu umgehen. Barber beschreibt die Knoten als rote und blaue Knoten für die beiden unterschiedlichen Typen. Dabei ist die Anzahl der roten Knoten gleich p und die Anzahl der blauen Knoten gleich q. Er definiert eine Modularitätsmatrix B_{ij}, die den Term $(A_{ij} - P_{ij})$ aus der Gleichung (1) enthält:

$$B_{ij} = A_{ij} - P_{ij} \tag{3}$$

Die Adjazenzmatrix A_{ij} für bipartite Netzwerke ist folgendermaßen definiert:

$$A = \begin{bmatrix} O_{p\times p} & \tilde{A}_{p\times q} \\ (\tilde{A}^{\intercal})_{q\times p} & O_{q\times q} \end{bmatrix} \tag{4}$$

$O_{i\times j}$ ist eine Nullmatrix mit i Zeilen und j Spalten. Durch die Umstellung kann das bipartite Netzwerk in einer Matrix repräsentiert werden.

Die Wahrscheinlichkeitsmatrix P, die das Nullmodell darstellt, wird folgendermaßen definiert:

$$P = \begin{bmatrix} O_{p\times p} & \tilde{P}_{p\times q} \\ (\tilde{P}^{\intercal})_{q\times p} & O_{q\times q} \end{bmatrix} \tag{5}$$

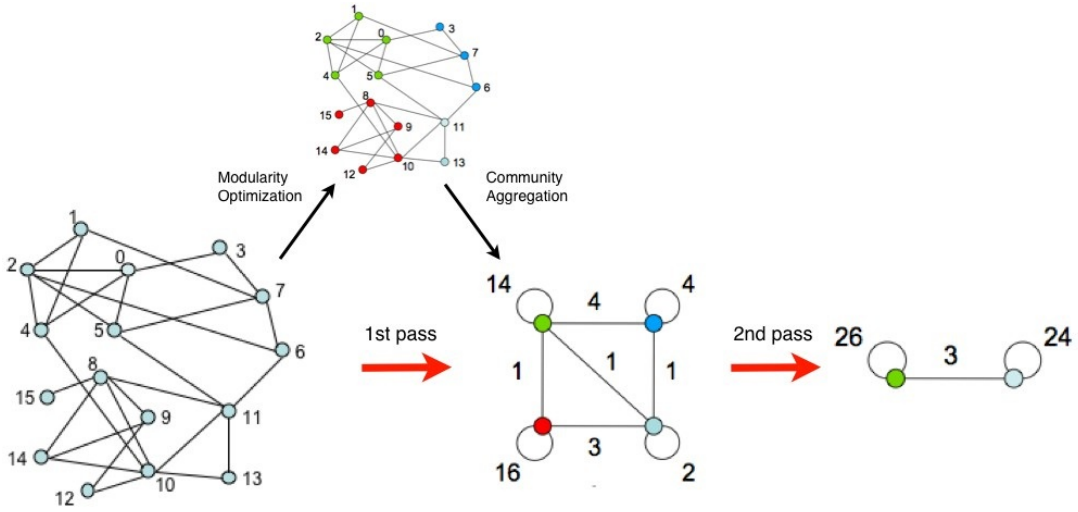
Wie in der Gleichung (3) kann $\tilde{B} = \tilde{A} - \tilde{P}$ gebildet werden und somit kann die Modularitätsmatrix B für bipartite Netzwerke folgendermaßen definiert werden:

$$B = \begin{bmatrix} O_{p\times p} & \tilde{B}_{p\times q} \\ (\tilde{B}^{\intercal})_{q\times p} & O_{q\times q} \end{bmatrix} \tag{6}$$

Mit dieser Umformung kann die Modularität für bipartite Netzwerke berechnet werden. Dies hat den Vorteil gegenüber (1), dass für gleichfarbige Knoten keine negativen Werte ausgerechnet werden, da gleichfarbige Knoten in diesem Fall den Wert 0 erhalten. Sei die Kantenanzahl der roten Knoten gleich k_i und die der blauen Knoten gleich d_j sowie m die Anzahl aller Kanten. Aus einer Reihe von Umformungen erhält man die Gleichung:

$$\tilde{P_{ij}} = \frac{k_i d_j}{m} \tag{7}$$

Somit kann äquivalent zur unipartiten Modularität auch eine Modularität für bipartite Graphen errechnet werden.



Quelle: Fast unfolding of communities in large networks (Blondel et al. , 2008)
https://arxiv.org/abs/0803.0476

Abbildung 1: Louvain-Algorithmus und seine zwei Durchgänge: Beim ersten Durchgang wird die Modularität durch lokale Änderungen der Communities optimiert. Beim zweiten Durchgang werden die neuen Communities zu einem Knoten zusammengeführt, um ein neues Netzwerk zu erstellen. Die Durchgänge werden solange wiederholt bis die Modularität nicht mehr erhöht werden kann.

4 Arzt-Patienten - Community Detection

In diesem Kapitel werden zwei bestehende Community Detection - Algorithmen beschrieben. Der erste Algorithmus, der *Louvain*-Algorithmus für unipartite Netzwerke, wird im Abschnitt 4.1 beschrieben. Der zweite Algorithmus, *LPAwb+*, wird im Abschnitt 4.2 beschrieben. Im letzten Abschnitt 4.3 wird der in dieser Arbeit entwickelte Algorithmus *Arztnetz-Detection* beschrieben, der aus einem gegebenen Arzt-Patienten-Netzwerk geeignete Arztnetze finden soll.

4.1 Louvain-Algorithmus

Der *Louvain*-Algorithmus zur Identifizierung von Communities in Netzwerken ist laut Blondel u.a. [Blo+08] einfach zu implementieren und für große Netzwerke geeignet. Er ist heuristisch und basiert auf der Optimierung der im Abschnitt 3.2 beschriebenen Modularität. Er besteht aus zwei Durchgängen (engl.: pass), die iterativ wiederholt werden. Initial erhält jeder Knoten seine eigene Community. In der Abbildung 1 sind die zwei Durchgänge bildlich dargestellt.

Im ersten Durchgang werden für jeden Knoten i seine Nachbarn j berücksichtigt. Der Knoten i wird aus seiner Community entfernt und der Community von j zugeordnet.

Der Zuwachs der Modularität für die neue Zuordnung wird berechnet. Die Community von j, bei dem der positive Zuwachs an Modularität für i am höchsten ist, wird die neue Gemeinschaft von i. Die erste Phase gilt als beendet, wenn dieser Schritt für alle Knoten i durchgeführt wurde. Mit dem ersten Durchgang ist das lokale Optimum der Modularität erreicht.

Im zweiten Durchgang des *Louvain*-Algorithmus wird jede Community als ein einziger Superknoten betrachtet. Die Verbindung zwischen zwei Communities ist die Summe des Gewichts der Verbindungen zwischen den Konten aus der ersten Community und der zweiten Community. Nach diesem Vorgehen ist das neue gewichtete Netzwerk und somit der zweite Durchgang komplett.

Nach der Vollendung des zweiten Durchgangs werden die zwei Durchgänge mit den neuen Superknoten wiederholt. Die Anzahl an kleinen Communities verringert sich mit jeder Iteration. Dieser Vorgang wird wiederholt bis die maximale Modularität erreicht ist. Die Iteration der zwei Durchgänge soll verhindern, dass die Verbesserung beim lokalen Optimum stecken bleibt.

Laut Blondel u.a. [Blo+08] bietet der Louvain-Algorithmus im Vergleich mit den Algorithmen von Clauset, Newman und Moore [CNM04], Pons und Latapy [PL05] und Wakita und Tsurumi [WT07] die besten Ergebnisse in Laufzeit und Modularität.

4.2 LPAwb+ Algorithmus

Der LPAwb+ Algorithmus von Beckett [Bec16] besteht wie der Louvain-Algorithmus auch aus zwei Schritten. Zur Einfachheit werden die zwei Knotentypen als rote Knoten r und blaue Knoten b bezeichnet. Initial bekommt jeder Knoten von der kleinsten der beiden Knotenmengen ein eindeutiges Label.

Im ersten Schritt wird die Label Propagation durchgeführt. Asynchron werden die blauen und die roten Knotenlabel aktualisiert, um ein lokales Optimum der Modularität zu erhalten. Label Propagation ermittelt initial für jeden Knoten ohne Label, welches Label aller Nachbarsknoten von der anderen Menge am meisten vorkommt. Dieses Label wird das neue Label für den Knoten.

Beckett schreibt außerdem weiterhin, dass für einen ausgewählten roten Knoten x ein neues Label g_x gewählt werden kann, indem folgende Bedingung maximiert wird:

$$g_x = (\sum_{v=1}^{b}(\tilde{W_{xv}} - \frac{y_x z_v}{M}))\delta(g, h_v) \tag{8}$$

Hierbei steht b für die blauen Knoten. Die gewichtete Adjazenzmatrix zwischen dem roten Knoten x und allen blauen Knoten v wird hier als $\tilde{W_{xy}}$ bezeichnet. Der gewichtete Knotengrad des roten Knoten x ist hier y_x und der von den blauen Knoten ist z_v. M ist die Summe aller Kantengraden. g ist das Label für die roten Knoten und h das Label für die blauen Knoten. Die Kronecker-Delta-Funktion $\delta(g, h_v)$ ist gleich 1, wenn beide Knoten das selbe Label haben und somit in derselben Community, ansonsten ist sie gleich 0.

Rote Knoten benutzen ausschließlich die Label (h) von den blauen Konten, um ihr eigenes Label (g) zu aktualisieren. Ähnlich benutzen die blauen Knoten ausschließlich die Labels der roten Knoten. Für die analogen Aktualisierungsregeln können folgende Gleichungen 9 und 10 definiert werden:

$$g_x\text{new} = \underset{g}{\text{argmax}}(N_{xg} - \frac{y_x Z_g}{M}) \tag{9}$$

$$h_x\text{new} = \underset{h}{\text{argmax}}(N_{xh} - \frac{Y_h z_x}{M}) \tag{10}$$

In der Gleichung 9 wird das neue Label $g_x new$ für den roten Knoten x jenes Label g, das die rechte Seite von der Gleichung 9 maximiert. Dabei ist N_{xg} die Summe von Verbindungen vom roten Knoten x zu anderen, die das Label g haben. Der Knotengrad y_x wird mit der Summe der Knotengrad Z_g aller mit x verbundenen Knoten mit dem Label g multipliziert und dann durch M dividiert. Analog verhält es sich mit der Aktualisierung des Labels der blauen Knoten durch die Betrachtung der Labels der roten Nachbarschaftknoten in der Gleichung 10. Die beiden Maximierungsgleichungen 9 und 10 werden asynchron ausgeführt. Zuerst werden die roten Knoten aktualisiert, dann die blauen, dann wieder die roten, solange bis die bipartite Modularität nicht mehr erhöht werden kann.

Der zweite Schritt funktioniert ähnlich wie der *Louvain*-Algorithmus. Nach dem ersten Schritt wird eine lokales Optimum der Modularität erreicht. Es werden kleine Communities miteinander verbunden, wenn sie die Modularität erhöhen und wenn keine anderen Communities vorhanden sind, mit denen eine Verbindung eine höhere Modularität herbeiführt.

Die Schritte 1 und 2 werden nun solange wiederholt bis keine Verbesserung der Modularität möglich ist.

Algorithm 1: Arztnetz-Detection

Input: Arzt-Patienten-Tabelle AP, Gütemaße B
Result: aktualisierte Arzt-Patienten-Tabelle AP, Arzt-Patienten-Tabelle mit Arztnetz-Zuordnung AP_{com}, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT

```
1  Initialisierung: flag := 1, AN := ∅, AT := ∅, AP_community := ∅
2  if (flag == 1 und Arzt-Patienten-Tabelle nicht leer) then
3  |  Phase 1 ;
4  |  Vergabe der Qualitätsklasse 1 in AP_com, AN, AT ;
5  end
6  if (flag == 2 und Arzt-Patienten-Tabelle nicht leer) then
7  |  Phase 2 ;
8  |  Vergabe der Qualitätsklasse 2 AP_com, AN, AT ;
9  end
10 if (flag == 3 und Arzt-Patienten-Tabelle nicht leer) then
11 |  Phase 3 ;
12 |  Vergabe der Qualitätsklasse 3 AP_com, AN, AT ;
13 end
14 if (flag == 4 und Arzt-Patienten-Tabelle nicht leer) then
15 |  Phase 4 ;
16 |  Vergabe der Qualitätsklasse 4 AP_com, AN, AT ;
17 end
18 return(AP, AP_com, AN, AT)
```

4.3 Arztnetz-Detection

Algorithmus 1 erhält als Eingabe die Arzt-Patienten-Tabelle AP sowie die expliziten Gütemaße für Arztnetze B. Er durchläuft vier Phasen. Aus jeder Phase können Communities identifiziert werden, die die Gütemaße B erfüllen. Geeignete Communities sind zugleich die Arztnetze. Jede Phase vergibt ihren Arztnetzen ihre Qualitätsklasse. Diese reichen von der Qualitätsklasse 1 aus der ersten Phase bis zur Qualitätsklasse 4 aus der vierten Phase.

In der ersten Phase (Seite 18) wird versucht, aus dem gewichteten bipartiten Netzwerk die Communities zu identifizieren. Dafür wird der Algorithmus *LPAwb+* von Beckett [Bec16] verwendet, der im Abschnitt 4.2 beschrieben wird.

In der zweiten Phase (Seite 19) wird das Arzt-Patienten-Netzwerk AP in ein Arzt-zu-Arzt-Netzwerk AA mit gemeinsamen Patienten als Verbindungen projiziert. Das Gewicht der Kanten im AA ist die Anzahl an gemeinsamen Patienten.

Bei diesem Schritt geht jedoch die Information über die Anzahl der Behandlungsfälle eines Patienten beim Arzt verloren, die in der ersten Phase noch berücksichtigt wurde. Für dieses gewichtete, unipartite Arzt-zu-Arzt-Netzwerk wird die Louvain-Methode aus dem Abschnitt 4.1 angewandt.

Die dritte Phase (Seite 21) bearbeitet initial große Communities, die aus der zweiten Phase entstanden sind und das Gütemaß *maximale Anzahl an Ärzten* überschreiten und das Gütemaß *minimale Anzahl an Fachgruppen* erfüllen.

Jede große Community wird nacheinander bearbeitet, indem die Anzahl an Ärzten ermittelt wird, die über der maximalen Anzahl n an Ärzten liegen. Daraufhin werden die n Ärzte ermittelt, die innerhalb der Community die wenigsten gemeinsamen Patienten hat. Diese werden aus der Community entfernt. Somit wird das Gütemaß *maximale Anzahl an Ärzten* erreicht. Da Ärzte weggeschnitten werden, müssen im Anschluss zwei Prüfungen stattfinden. Die erste Prüfung betrifft das Gütemaß *minimale Anzahl an Fachgruppen* und die zweite Prüfung prüft, ob die verbliebenen Ärzte im AA noch als Komponente ausgeführt werden. Nach den zwei erfolgreichen Prüfungen können diese Communities abgespeichert werden.

Die vierte Phase (Seite 23) wird lediglich aktiviert, wenn nach dem Durchlauf aller Phasen noch Ärzte übrigbleiben, die keinem Arztnetz zugeordnet werden konnten, und wenn lediglich zu kleine Communities (nach der zweiten oder nach der dritten Phase) existieren. Hier wird wiederum ein gewichtetes, unipartites Arzt-zu-Arzt-Netzwerk AA aus den übriggebliebenen Ärzten betrachtet. Jedoch werden hier keine Community Detection - Methoden mehr angewandt, sondern die Komponenten als Communities angenommen und die Gütemaße geprüft. Bei Erfüllung aller Gütemaße werden die entsprechenden Communities abgespeichert. Nach diesem Schritt können Ärzte ohne Arztnetze übrigbleiben.

Zum Schluss werden vier Tabellen ausgegeben. Die erste Tabelle ist die Arzt-Patienten-Tabelle AP, die verkleinert wird, wenn geeignete Communities gefunden werden. Die aktuelle Tabelle AP kann auch leer sein, wenn jeder Arzt einem Arztnetz zugeordnet werden kann. Die zweite Tabelle ist AP_{com}. Diese enthält die Arzt-Patienten-Verbindungen aus AP sowie die jeweilige Arztnetz-ID für den Arzt. Die dritte Tabelle ist die Arztnetz-Tabelle AN. Diese enthält Informationen über die Arztnetze wie die Anzahl an Ärzten, Patienten und andere Informationen. Die vierte resultierende Tabelle ist die Arzttabelle mit Arztnetz-Zuordnung AT, die die Ärzte mit ihrer Fachgruppe und Arztnetzzuordnung auflistet.

In den folgenden Abschnitten 4.3.1, 4.3.2, 4.3.3 und 4.3.4 werden die vier Phasen detaillierter beschrieben.

Der R-Code für den oben genannten Algorithmus kann auf dieser Seite gefunden werden: https://github.com/frankzng/arztnetz_detection.

Algorithm 2: Phase 1

Input: Arzt-Patienten-Tabelle AP, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT, Gütemaße B

Result: aktualisierte Arzt-Patienten-Tabelle AP, Arzt-Patienten-Tabelle mit Arztnetz-Zuordnung AP_{com}, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT

```
do
    LPAwb+ Algorithmus auf AP;
    gefundene Communities auf Gütemaße B prüfen;
    if geeignete Communities gefunden then
        AN, AT, AP_com aktualisieren (erweitern) ;
        AP aktualisieren (verkleinern) ;
    else
        flag := 2
    end
while (flag == 1 und geeignete Communities gefunden und Arzt-Patienten-Tabelle nicht leer);
return(AP, AP_com, AN, AT)
```

4.3.1 Phase 1: Bipartite Community Detection

In **Phase 1** wird der Algorithmus 2 verwendet. Hier wird aus der Arzt-Patienten-Tabelle AP ein gewichtetes bipartites Arzt-Patienten-Netzwerk erstellt und mit dem *LPAwb+ Algorithmus* in Communities partitioniert. Jede Community wird dann auf die expliziten Gütemaße geprüft. Werden die Gütemaße bei einer Community erfüllt, gilt diese Community als *geeignete Community*, nun auch als *Arztnetz* bezeichnet.

Die Arztnetz-Tabelle AN wird um die gefundenen Arztnetze erweitert und die dazugehörenden Ärzte werden in der Arzttabelle mit Arztnetz-Zuordnung AT gespeichert. Die Arzt-Patienten-Tabelle AP wird um die Arzt-Patienten-Verbindungen verkleinert, wenn sie die Ärzte mit Arztnetzzuordnungen enthält.

Daraufhin wird aus der aktuellen, kleineren Tabelle AP ein neues gewichtetes, bipartites Netzwerk erstellt und der *LPAwb+ Algorithmus* erneut durchgeführt. Alle entsprechenden Tabellen werden bei erfolgreicher Suche und nach erfolgreichen Prüfungen wie im Vorschritt aktualisiert. Alle Arztnetze aus der Phase erhalten die Qualitätsklasse 1.

Dieser Prozess wird solange durchgeführt bis keine geeigneten Communities mehr gefunden werden. Der Flag wird auf 2 gesetzt, um die Phase 2 (Seite 19) einzuleiten. Die Phase 1 ist damit beendet.

Algorithm 3: Phase 2

Input: Arzt-Patienten-Tabelle AP, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT, Gütemaße B

Result: `// APGC ist eine Liste von großen Communities, wenn große Communities gefunden werden und Phase 3 eingeleitet werden soll ;`

```
1  Wenn Phase 3 eingeleitet wird: APGC, AP, AP_com, AN, AT
2  Wenn Phase 4 eingeleitet wird: AP, AP_com, AN, AT ;
3  do
4      Aus AP ein unipartites Arzt-zu-Arzt Netzwerk AA mit gemeinsamen Patienten als
        Kantengewicht erstellen ;
5      Louvain-Methode auf AA;
6      gefundene Communities auf Gütemaße B prüfen;
7      if geeignete Communities gefunden then
8          AN, AT, AP_com aktualisieren (erweitern) ;
9          AP aktualisieren (verkleinern) ;
10         result := (AP, AP_com, AN, AT)
11     end
12     if keine geeigneten Communities gefunden, aber große Communities gefunden then
13         Erstellung Liste von großen Communities APGC ;
14         AP verkleinern; flag := 3 ;
15         result := (APGC, AP, AP_com, AN, AT)
16     end
17     if keine geeigneten Communities gefunden, aber kleine Communities gefunden then
18         flag := 4 ;
19         result := (AP, AP_com, AN, AT)
20     end
21 while (flag == 2 und geeignete Communities gefunden und Arzt-Patienten-Tabelle nicht
    leer);
22 return(result)
```

4.3.2 Phase 2: Unipartite Community Detection

In **Phase 2** wird der Algorithmus 3 verwendet. Aus der Arzt-Patienten-Tabelle AP wird statt des bipartiten nun ein unipartites Arzt-zu-Arzt-Netzwerk AA erstellt. Das Gewicht der Kanten ist die Anzahl gemeinsamer Patienten. Die Anzahl der Behandlungsfälle, die noch als Gewicht der Kanten zwischen Arzt und Patient dient, wird nicht mehr berücksichtigt. Hier findet eine Projektion des bipartiten Netzwerkes in ein unipartites Netzwerk mit neuem Kantengewicht statt.

Nach der Projektion wird auf das Netzwerk AA die Louvain-Methode angewendet. Die identifizierten Communities werden auf die Gütemaße B überprüft. Geeignete Gemeinschaften gelten als Arztnetze der Qualitätsklasse 2. Die Arzt-Patienten-Tabelle AP wird um die Ärzte verkleinert, die einem Arztnetz zugeordnet werden. Die Arzttabelle AT wird um diese Ärzte wiederum erweitert.

Phase 2 wird mit der aktualisierten AP solange wiederholt bis keine Arztnetze mehr gefunden werden. In dem Falle, dass keine Arztnetze mehr gefunden werden, die beste-

henden Communities aber zu groß sind, werden diese in einer Extra-Tabelle *APGC* bestehend aus Verbindungen zwischen Ärzten und Patienten sowie der Community-Nummern, gespeichert. Entsprechend wird die Tabelle *AP* verkleinert und das $flag := 3$ gesetzt. Die Phase 2 würde hier mit der Herausgabe der Tabelle *APGC* enden, die weiter in Phase 3 (Seite 21) bearbeitet wird.

Alle Arztnetze aus dieser Phase erhalten die Qualitätsklasse 2.

Werden weder geeignete noch große Communities identifiziert, wird das $flag := 4$ gesetzt und folglich mit Phase 4 (Seite 23) fortgesetzt.

Algorithm 4: Phase 3

Data: Arzt-Patienten-Tabelle AP, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT, Liste von großen Communities $APGC$, Gütemaße B
Result: AP, AP_{com} mit der Arztnetz-Zuordnung, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT

```
1  do
2      do
3          for Community c in APGC do
4              Verbindungen aus AP mit Ärzten aus Community c in
                Arzt-zu-Arzt-Netzwerk AA projizieren ;
5              Louvain-Methode auf AA ;
6              if geeignete Communities gefunden then
7                  AN, AT, AP_com aktualisieren (erweitern) ;
8                  AP aktualisieren (vergrößern) // Hier werden die
                       rausgeschnittenen Ärzte mit ihren Patienten-Verbindungen
                       wieder in AP importiert
9                  flag := 2 ;
10             else
11                 AP aktualisieren (vergrößern) // Hier werden die alle Ärzte aus
                       Community c mit ihren Patienten-Verbindungen wieder in AP
                       importiert
12                 flag := 4 ;
13             end
14         end
15     while (flag == 3 und geeignete Communities gefunden und Arzt-Patienten-Tabelle
         nicht leer);
16     activate(Phase 2)
17 while (flag == 3 und geeignete Communities gefunden und Arzt-Patienten-Tabelle nicht
     leer);
18 return(AP, AP_com, AN, AT)
```

4.3.3 Phase 3: Große Communities

In **Phase 3** wird Algorithmus 4 aktiviert. Jede zu große Community c in $APGC$ bearbeitet. In jeder großen Community wird ermittelt, wie viele Ärzte raus aus der Community entfernt werden müssen, bis die Größe der Gemeinschaft die maximal erlaubte Anzahl an Ärzten enhält (zu entfernende Anzahl = n). Im folgenden Schritt werden n Ärzte mit den wenigsten gemeinsamen Patienten aus der Community entfernt. Die verkleinerte Community wird auf das Gütemaß *minimale Anzahl an Fachgruppen* geprüft und ob sie noch eine Komponente ist. Treffen beide Bedingungen zu, wird diese als Arztnetz der Qualitätsklasse 3 abgespeichert.

Nachdem alle Communities c abgearbeitet wurden, werden die entfernten Ärzte der Tabelle AP hinzugefügt. Aus AP wird wieder das unipartite Arzt-zu-Arzt-Netzwerk AA erstellt, auf welchem der *Louvain*-Algorithmus wie in Phase 2 durchgeführt wird. Die Communities werden wiederum auf die Gütemaße B geprüft. Geeignete Communities werden als Arztnetze der Qualitätsklasse 3 abgespeichert, zu große Gemeinschaften wer-

den wie oben beschrieben durch Entfernung der schwächsten Ärzte mit anschließender Überprüfung bearbeitet.

Wenn alle Communities c abgearbeitet wurden und mindestens eine Community c als geeignet gespeichert wurde, wird die der Algorithmus 3 der Phase 2 wieder aktiviert. Communities, die jedoch in dieser späten Phase 2 gefunden werden, erhalten die Qualitätsklasse 3. Die Phase 2 vergibt entweder $flag := 3$, wenn zu große Communities gefunden werden, oder $flag := 4$, wenn zu kleine Communities gefunden werden.

Dies wird solange durchgeführt bis entweder keine Ärzte mehr ohne Arztnetz oder nur noch zu kleine Communities vorhanden sind. Falls noch zu kleine Communities vorhanden sind und die Anzahl an Ärzten insgesamt noch für mindestens ein Arztnetz ausreicht, wird Phase 4 aktiviert.

Algorithm 5: Phase 4

Data: Arzt-Patienten-Tabelle AP, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT, Gütemaße B

Result: AP, AP_{com} mit der Arztnetz-Zuordnung, Arztnetz-Tabelle AN, Arzttabelle mit Arztnetz-Zuordnung AT

```
1 Projektion in unipartites Arzt-zu-Arzt Netzwerk AA ;
2 Komponenten im AA identifizieren ;
3 Komponenten auf Gütemaße B prüfen ;
4 if geeignete Communities bzw. Komponenten gefunden then
5 |   AN, AT, AP_com aktualisieren (erweitern) ;
6 |   AP aktualisieren (verkleinern) ;
7 end
8 return(AP, AP_com, AN, AT)
```

4.3.4 Phase 4: Kleine Communities

In **Phase 4** wird Algorithmus 5 aktiviert. Dieser projiziert die AP-Tabelle in ein Arzt-zu-Arzt-Netzwerk AA. Aus diesem werden die Komponenten k identifiziert. Die Komponenten k werden als Communities betrachtet und werden auf die Gütemaße überprüft. Communities, die alle Gütemaße erfüllen, werden als Arztnetze der Qualitätsklasse 4 abgespeichert. Ärzte, die nach dieser Phase keinem Arztnetz gehören, werden zusammen mit ihren Patienten in AP abgespeichert.

Diese Phase ist die letzte des Algorithmus 1 und die Ergebnisse spätestens nach dieser Phase sind die Endergebnisse.

5 Experiment und Evaluierung

Im Abschnitt 5.1 werden die Hard- und Software beschrieben, die für die Analysen genutzt werden. Im Abschnitt 5.2 wird der Aufbau der Testdaten erklärt. Der interessante Teil ist im Abschnitt 5.3. In diesem werden die Ergebnisse bewertet. Im Abschnitt 5.4 werden die Probleme und Herausforderungen beschrieben, die nach den Tests ersichtlich werden.

5.1 Hardware und Software

Durchgeführt wird der Algorithmus auf einem Rechner mit Intel Xeon CPU E5-2620 v4 mit 2.10 GHz. Der Algorithmus läuft auf einem Kern.

Der Algorithmus ist in der Sprache R-Statistics in der Version 3.3.3 geschrieben. Verwendet werden die Pakete *bipartite*, *dplyr*, *igraph* und *Matrix*. *dplyr* und *Matrix* werden für die Manipulation von Tabellen und Aufbau von Matrizen verwendet.

bipartite enthält die Funktion *compute_modules*, die ein gewichtetes bipartites Netzwerk clustert. Hier wird die Methode *beckett* schon per Default in *compute_modules* angewendet. Per Default wird die Option *DIRTLPAwb+* durchgeführt. Dabei wird *LPAwb+* mehrmals durchgeführt. Darauf wird in dieser Arbeit auf Grund der Netzwerkgrößen und der langen Laufzeit verzichtet. Daher wird die Option in *forceLPA* geändert, um *LPAwb+* für einen einzigen Lauf durchzuführen.

In *igraph* ist die Methode *cluster_louvain* vorhanden. Diese wird für Phase 2 und Phase 3 verwendet. Außerdem bietet *igraph* Funktionen, um Nachbarschaftstabellen in *igraph*-Objekte zu überführen, die Graphen repräsentierten.

5.2 Testdaten

Aus den vertragsärztlichen Abrechnungsdaten werden Patienten ausgewählt, die im 4. Quartal 2016 von mindestens 5 verschiedenen Ärzten behandelt wurden. Arztnetze sind für Patienten mit einem hohen Versorgungsbedarf interessant. Patienten, die keinen regelmäßigen Versorgungsbedarf aufweisen, könnten die Ergebnisse verzerren, da sie eher zu einer zufälligen Arztwahl neigen. Die Patienten sind nach ihren Wohnkreisen unterteilt.

Die behandelnden Ärzte werden diesen Patienten zugeordnet. Für die Ärzte werden in den Abrechnungsdaten keine Kreisdaten ausgewiesen. Lediglich ihre KV-Zugehörigkeit ist vorhanden, welche mehrere Kreise umfassen kann. Das bedeutet, dass ein Arzt in mehreren Netzwerken auftauchen könnte. Die Fachgruppe des Arztes wird mit angegeben. Für jeden Kreis wird somit eine Arzt-Patienten-Tabelle bestehend aus der Identifikation-Nummer des Arztes sowie des Patienten, Fachgruppe des Arztes und die Anzahl der Behandlungsfälle erstellt. Auf diese Testdaten konnten in der Datenstelle beim Zentralinstitut der kassenärztlichen Versorgung in Deutschland für wissenschaftliche Zwecke zugegriffen werden. Da es sich um Sozialdaten aus dem System der gesetzlichen Krankenversicherung handelt, stehen sie nicht öffentlich zur Verfügung. Die Ergebnisse sind aggregiert und können nicht auf einzelne Ärzte oder Patienten zurückgeführt werden.

5.3 Ergebnisse

In der Tabelle 1 werden einige deskriptive Zahlen aufgelistet:

203 Kreise werden getestet. Im Durchschnitt behandeln 154 Ärzte verschiedener Fachgruppen 5.130 Patienten mit einem hohen Versorgungsbedarf. Für diese Kreise können circa 3 bis 4 Arztnetze im Durchschnitt identifiziert werden. Diesen werden im Schnitt 150 Ärzte zugeordnet. Das bedeutet, dass lediglich 4 Ärzte pro Kreis keinem Arztnetz zugeordnet werden.

In jedem Arztnetz behandeln im Schnitt 42 Ärzte durchschnittlich 2.674 Patienten. Dabei werden lediglich 18,5% von diesen Patienten in nur einem Arztnetz behandelt.

Die Dichte und der Cluster-Koeffizient werden über das unipartite Arzt-zu-Arzt-Netzwerk berechnet. Das Kantengewicht, in diesem Fall die Anzahl gemeinsamer Patienten, spielt hier keine Rolle. Die Dichte muss hier kritisch betrachtet werden, da sie unabhängig von der Anzahl der gemeinsamen Patienten ist. Wenn drei Ärzte sich je einen Patienten teilen, dann ist die Dichte für diese drei Ärzte genauso hoch wie die Dichte von drei Ärzten, die sich 1.000 Patienten teilen.

Die Dichte eines Arztnetzes ist im Schnitt 0,74 und somit um 35% höher als die durchschnittliche Dichte von 0,55 in einem Kreis. Auch der durchschnittliche Cluster-Koeffizient pro Arztnetz ist mit 0,83 um 15% höher als der in einem durchschnittlichen Kreis.

In den weiteren Abschnitten werden weitere Kennzahlen genauer betrachtet.

Tabelle 1: Tabelle mit Daten für Kreise und Arztnetze

Deskriptive Statistiken	
Anzahl Kreise	203
Durchschnittliche Anzahl an Arztnetzen pro Kreis	3,65
Minimale Anzahl an Ärzten in einem Kreis	43
Maximale Anzahl an Ärzten in einem Kreis	337
Durchschnittliche Anzahl an Ärzten in einem Kreis	154
Durchschnittliche Anzahl an Ärzten in einem Arztnetz	42
Durchschnittliche Anzahl an Ärzten in einem Kreis mit Arztnetz	150
Minimale Anzahl an Patienten in einem Kreis	336
Maximale Anzahl an Patienten in einem Kreis	14.414
Durchschnittliche Anzahl an Patienten in einem Kreis	5.130
Durchschnittliche Anzahl an Patienten in einem Arztnetz	2.674
Minimale Anzahl an Fällen in einem Kreis	2.578
Maximale Anzahl an Fällen in einem Kreis	167.734
Durchschnittliche Anzahl an Fällen in einem Kreis	50.294
Durchschnittliche Anzahl an Fällen in einem Arztnetz	13.861
Minimale Dichte eines Kreises	0,23
Maximale Dichte eines Kreises	0,93
Durchschnittlichte Dichte eines Kreises	0,55
Minimale Durchschnitts-Dichte eines Arztnetzes	0,50
Maximale Durchschnitts-Dichte eines Arztnetzes	0,94
Durchschnittlichte Dichte eines Arztnetzes	0,74
Minimaler Cluster-Koeffizient eines Kreises	0,51
Maximaler Cluster-Koeffizient eines Kreises	0,94
Durchschnittlichter Cluster-Koeffizient eines Kreises	0,72
Minimaler Durchschnitts-Cluster-Koeffizient eines Arztnetzes	0,65
Maximaler Durchschnitts-Cluster-Koeffizient eines Arztnetzes	0,95
Durchschnittlichter Cluster-Koeffizient eines Arztnetzes	0,83
Minimaler Anteil von exklusiven Patienten in einem Arztnetz	0,013
Maximaler Anteil von exklusiven Patienten in einem Arztnetz	1,000
Durchschnittlichter Anteil von exklusiven Patienten in einem Arztnetz	0,185

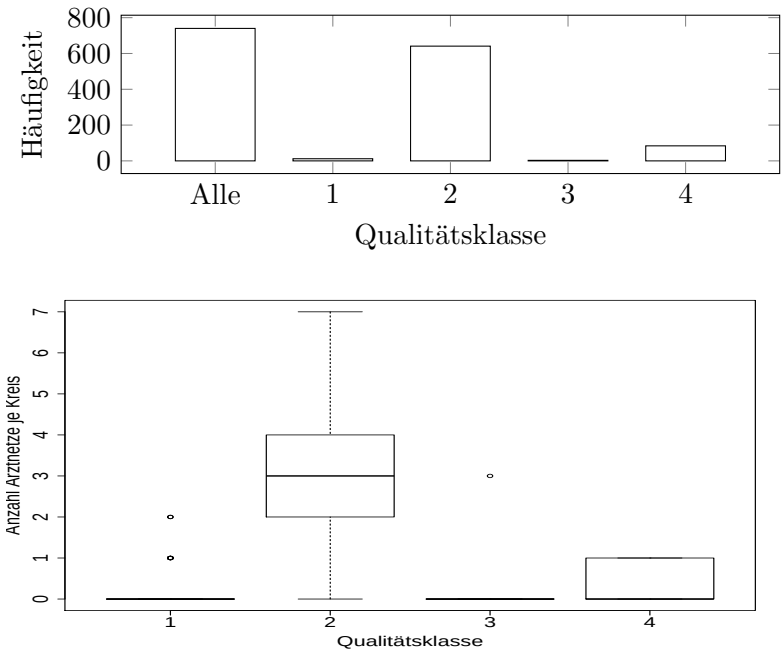


Abbildung 2: *Oben:* Anzahl der Arztnetze je Qualitätsklasse bei der Analyse von 203 Kreisen mit den Gütemaßen: 20 bis 100 Ärzte und mindestens 3 verschiedene Fachgruppen. *Unten:* Durchschnittliche Anzahl der Arztnetze pro Kreis, nach Qualitätsklasse unterteilt.

5.3.1 Anzahl der Arztnetze verschiedener Qualitätsklassen

In der Abbildung 2 ist ersichtlich, dass für die Gütemaße 20 bis 100 Ärzte und mindestens 3 Fachgruppen in einem Arztnetz fast keine Arztnetze der Qualitätsklasse 1 (aus der Phase 1) und fast keine Arztnetze der Qualitätsklasse 3 (aus der Phase 3) gefunden werden. Insgesamt werden 12 Arztnetze der Qualitätsklasse 1 für 203 Kreise und 3 Arztnetze der Qualitätsklasse 3 von insgesamt 740 Arztnetzen über alle Kreise gefunden.

Die meisten Arztnetze werden in Phase 2 gefunden, wenn das Netzwerk in ein unipartites Netzwerk projiziert und der *Louvain*-Algorithmus angewandt wird. Über 86,6% der Arztnetze (in absoluter Zahl: 641) haben die Qualitätsklasse 2. Von der Qualitätsklasse 4, also Arztnetze aus zu kleinen Communities werden 84 entdeckt.

Zu große Communities sind eher selten, da die Gütemaße mit bis zu 100 Ärzten zu groß zu sein scheint, um solche Communities für Phase 3 zu erhalten. Diese Zahl könnte sich bei viel größeren Kreisen ändern, weil die Gütemaße wohl eher einzuhalten sind.

In der unteren Abbildung ist zu sehen, dass in den meisten Kreisen die Anzahl der Arztnetze zwischen 2 und 4 beträgt. Ein Kreis hat zwei Arztnetze der Qualitätsklasse 1, zehn der 203 Kreise haben jeweils ein Arztnetz der Klasse 1. Lediglich der größte Kreis hat ein Arztnetz der Qualitätsklasse 3.

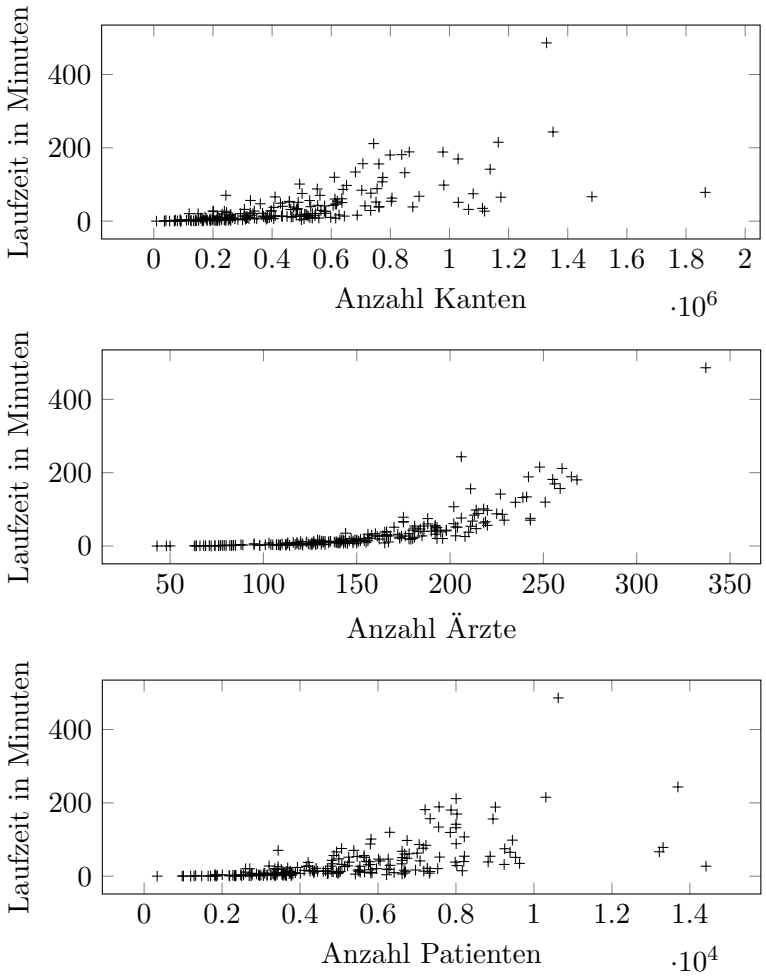


Abbildung 3: Die Laufzeiten abhängig von drei verschiedenen Parametern.

5.3.2 Laufzeit

in der Abbildung 3 werden die Laufzeiten in Minuten für die Anzahl an Kanten, Anzahl an Ärzten und Anzahl an Patienten verdeutlicht. Hier ist gut zu erkennen, dass die Anzahl an Ärzten die Laufzeit sehr beeinflusst.

Die Vermutung, dass die Anzahl an Kanten eher Einfluss auf die Laufzeit nimmt, bestätigt sich, zumindest anhand der Bilder nicht. Das kann daran liegen, dass in der ersten Phase der *LPAwb+*-Algorithmus verwendet wird, der höchstens so viele Communities erstellen kann wie die Anzahl der kleinsten Menge ist. Da die Arztmenge viel kleiner ist als die Patientenmenge, bestimmt diese sozusagen die Anzahl an Communities, die im Prozess der Optimierung der Modularität von Barber immer wieder *geändert* werden müssen. Bemerkenswert ist, dass der dichteste Kreis mit $> 1.8 \cdot 10^6$ Kanten, weniger Zeit in Anspruch nimmt als manche Kreise mit weniger Kanten.

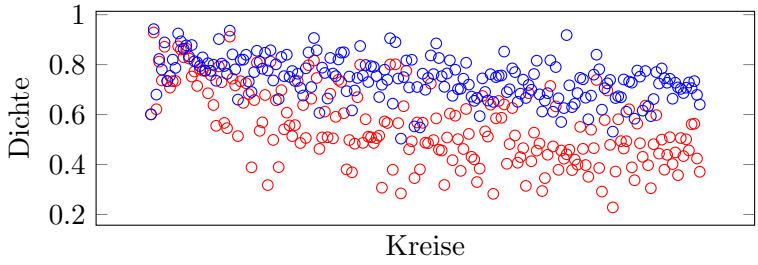


Abbildung 4: In dieser Abbildungen stellen die roten Kreise die Dichte pro Kreis dar. Die blauen Kreise stellen die Durchschnittsdichte der Arztnetze pro Kreis dar.

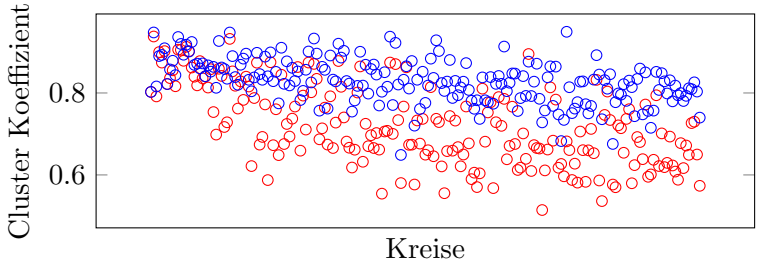


Abbildung 5: In dieser Abbildungen stellen die roten Kreise den Cluster-Koeffizient pro Kreis dar. Die blauen Kreise stellen der durchschnittlichen Cluster-Koeffizient der Arztnetze pro Kreis dar.

5.3.3 Dichte und Cluster-Koeffizient

In den Abbildungen 4 und 5 auf der Seite 29 wird die Dichte und der Cluster-Koeffizient je Kreis mit der durchschnittlichen Dichte sowie dem durchschnittlichen Cluster-Koeffizienten von allen Arztnetzen je Kreis verglichen. Die roten Kreise stellen die Dichte, beziehungsweise den Cluster-Koeffizienten der Kreise dar. Die blauen Kreise stellen die Dichte, beziehungsweise den Cluster-Koeffizienten der Arztnetze je Kreis dar.

In diesen Abbildungen wird deutlich, dass die produzierten Arztnetze allgemein dichter und der Cluster-Koeffizient auch höher als in den nichtpartitionierten Kreisen.

Diese Eigenschaften entsprechen den weichen Kriterien in den Gütemaßen. Somit erfüllt der Algorithmus neben der Einhaltung der harten Kriterien auch den Zweck, dichtere Gruppen von Ärzten zu finden.

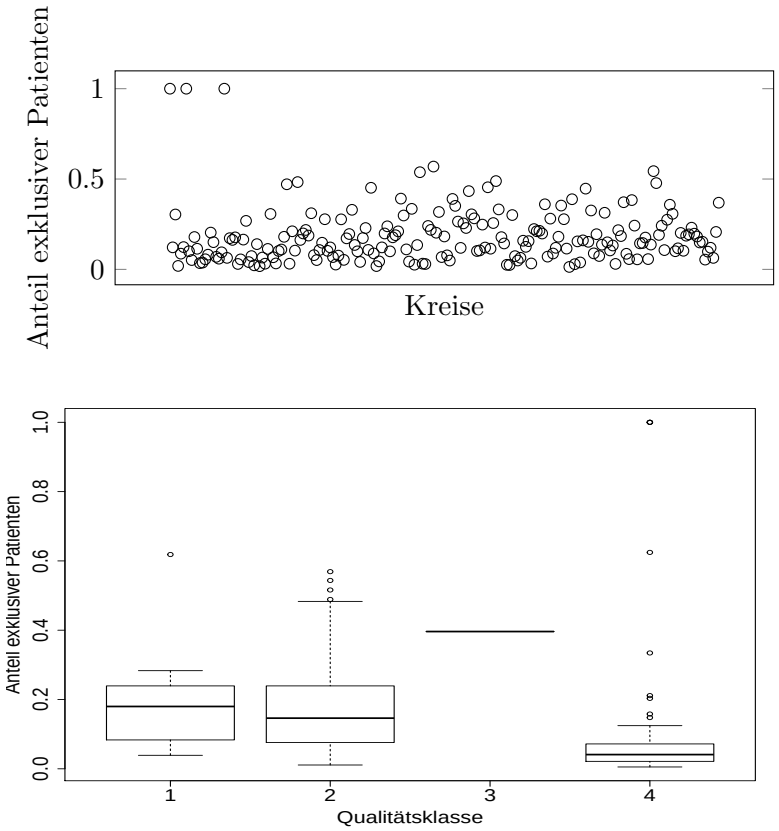


Abbildung 6: *Oben:* Hier wird für jeden Kreis der Anteil an exklusiven Patienten dargestellt. *Unten:* Der Anteil an exklusiven Patienten wird für jede Qualitätsklasse dargestellt.

5.3.4 Exklusive Patienten

Eine wichtige Eigenschaft von Arztnetzen ist, dass die Patienten möglichst nur in ein Arztnetz eingebunden sind. In den Abbildungen 6 sieht man einerseits für alle Kreise, wie der durchschnittliche Anteil an exklusiven Patienten pro Kreis ist. Dabei wird festgestellt, dass in den meisten Kreisen die Arztnetze lediglich bis zu 25 Patienten pro 100 Patienten exklusiv nur das eine Arztnetz besuchen.

In der unteren Abbildung wird der durchschnittliche Anteil an exklusiven Patienten für jede Qualitätsklasse betrachtet. Das Ergebnis zeigt, dass der Anteil an exklusiven Patienten für Arztnetze der Qualitätsklasse 1 im Durchschnitt etwas höher liegt als der Anteil von den Arztnetzen der Qualitätsklasse 2. Die Arztnetze der Qualitätsklasse 3 liegen bei circa 40%. Hier muss jedoch beachtet werden, dass die Anzahl an Arztnetzen mit 12 bzw. 3 für die Qualitätsklassen 1 bzw. 3 ziemlich gering sind. Die Arztnetze der Qualitätsklasse 4 haben deutlich weniger exklusive Patienten als die höheren Qualitätsklassen.

5.4 Probleme und Herausforderung

Bei der Entwicklung und beim Testen des Algorithmus sind einige Punkte aufgefallen, die für die Zukunft zu beachten sind.

Es existieren derzeit einige Community Detection-Methoden, die für die Identifizierung von Arztnetzen mit festen Gütemaßen geeigneter sein können.

Auch die Verwendung der bipartiten Community Detection *LPAwb+* muss hinterfragt werden, da diese für die Testnetzwerke fast keine Arztnetze gefunden hat. Im Abschnitt 5.4.1 wird diese Problematik erläutert.

Eine weitere große Herausforderung ist die Auswahl der Gütemaße. Für diese Arbeit wurden relativ einfache Kriterien verwendet, die leicht zu überprüfen sind. Jedoch stellt sich hier auch die Frage, ob diese ausreichen oder ob nicht eventuell weitere Gütemaße entwickelt werden müssten. Dies wird was im Abschnitt 5.4.2 ausführlicher erläutert.

5.4.1 Barbers Modularität: Problem für Arztnetz-Detection

Die Besonderheit in realen Arzt-Patienten-Netzwerken ist, dass die Anzahl an Patienten die Anzahl an Ärzten bei weitem übersteigt. So können auf einen Arzt mehr als 1000 Patienten zugeordnet werden.

Aus dem Test geht hervor, dass lediglich $< 1\%$ der Arztnetze aus der Phase 1 entstehen. Somit stellt sich die Frage, ob die Optimierung der Modularität von Barber zumindest für Arztnetze überhaupt geeignet ist.

Da die Modularitätsmatrix wie im Abschnitt 3.3 beschrieben nun die folgende Formel enthält: $B_{ij} = A_{ij} - P_{ij}$. A_{ij} ist nur dann größer 0, wenn der Arzt i und der Patient j eine Verbindung haben und davon jedoch $\tilde{P_{ij}} = \frac{k_i d_j}{m}$ abgezogen wird. k_i und d_j beschreiben die Anzahl an Verbindungen einerseits des Arztes zu seinen Patienten und andererseits die Anzahl an Verbindungen des Patienten zu seinen Ärzten. Daraus ist ersichtlich, dass $\tilde{P_{ij}}$ ziemlich groß wird, wenn die Patienten willkürlich zu vielen verschiedenen Ärzten gehen. Das bedeutet, wenn viele Ärzte in einer Community auftauchen und diese Ärzte neben vielen gemeinsamen *exklusiven* Patienten auch Patienten haben, die auch anderen Communities zugeordnet werden können, dann wird $\tilde{P_{ij}}$ so groß, dass die Modularität abnimmt. Die Modularität kann erhöht werden, wenn Ärzte aus dieser Community entfernt werden. Und der Anteil an exklusiven Patienten nimmt zu.

Für die Zukunft müsste dieser Umstand berücksichtigt werden. Eine Idee ist ein weiteres Gewicht einzuführen, sodass Verbindungen von Patienten zu anderen Communities weniger bestraft werden. Damit lassen sich mehr Arztnetze aus der Phase 1 generieren.

5.4.2 Weitere explizite Gütemaße

In dieser Arbeit werden die drei wichtigsten Gütemaße verwendet. Jedoch ist es unwahrscheinlich, dass die Anzahl an Ärzten sowie die Anzahl an verschiedenen Fachgruppen ausreichend für die Erstellung von Arztnetzen in der Praxis.

Für eine Anwendung in der Praxis sollte mit Fachleuten besprochen werden, welche Kriterien für Arztnetze noch zu erfüllen beziehungsweise wünschenswert sind, um die aus dem Algorithmus entstandenen Arztnetze in der Praxis zu verwirklichen.

Ein mögliches weiteres Kriterium kann die Anzahl der Fachgruppen, sodass die Arztnetze möglichst viele Fachgruppen enthalten. Dies würde zwar weniger gut verbundene Ärzte in ein Arztnetz zusammenfassen, dafür aber die Verbindungen der verschiedenen Fachgruppen stärken. Möglicherweise werden dadurch die Dichte oder der Cluster-Koeffizient verschlechtert. Für die Qualität der Arztnetze hinsichtlich der praktischen Zusammenarbeit von Ärzten dürfe diese Art von Regulierung einen Vorteil darstellen.

6 Fazit

In dieser Arbeit wurde ein Algorithmus entwickelt, der mit Hilfe von bestehenden Community Detection-Algorithmen reale Arzt-Patienten-Netzwerke in Arztnetze aufteilt, wobei vordefinierte Gütemaße beachtet werden.

Die Laufzeit steigt exponentiell mit der Anzahl an Ärzten, was zuerst verwundert. Dies widerspricht zunächst der Annahme, dass die Anzahl an Kanten die Laufzeit am größten beeinflussen würde.

Bei durchschnittlich 154 Ärzte in jedem Kreis konnten für circa 149 Ärzte ein Arztnetz zugeordnet werden. Dabei ist die durchschnittliche Dichte und der durchschnittliche Cluster-Koeffizient der Arztnetze durchweg höher, als die der Netzwerke ohne die Aufteilung in Arztnetze.

Die Ergebnisse der Tests zeigen aber auch, dass der Algorithmus noch zu überarbeiten ist. Viele Arztnetze entstehen aus der zweiten Phase. Die erste Phase mit der Nutzung der bipartiten Struktur und des Algorithmus für die Identifizierung von Communities in bipartiten Netzwerken ist sehr Laufzeit-intensiv, bringt jedoch fast keine Ergebnisse. Das macht die erste Phase obsolet.

Nichtsdestotrotz zeigt sich, dass ein eigener Algorithmus für eine spezifische Aufteilung mit Hilfe von Community Detection - Methoden realisierbar und nützlich ist.

Ausblick

Für den realen Einsatz des Algorithmus besteht noch Bearbeitungsbedarf besonders in der ersten Phase des Algorithmus. Das diese Phase beim Testen die meiste Zeit einnimmt und in Hinsicht auf die Arztnetze wenige Ergebnisse bringen, muss die Modularität überdacht werden.

Es kann jedoch auch von der Modularität abgesehen werden, da das Ziel eine dichte Gruppierung von Ärzten ist. Dies kann mit Experten in diesem Bereich zusammen mit den realen Voraussetzungen für ein Arztnetz besprochen werden.

Die Laufzeit ist zwar für die ersten 203 Kreise noch akzeptabel. Die längste Laufzeit betrugt circa 486 Minuten für 337 Ärzte und 10624 Patienten. Circa 8 Stunden sind zwar noch praxistauglich, es stellt sich dennoch die Frage, wie lange der Algorithmus für die weiteren 200 größeren Kreise braucht. Vor allem wären Kreise wie Berlin oder München Laufzeit-technisch eine zu große Herausforderung. Dies sollte optimiert werden, da vor allem für die großen Kreise beziehungsweise Städte eine Zusammenarbeit von Ärzten erwünscht.

Literatur

[Bar+12] M. L. Barnett, N. A. Christakis, J. O'Malley, J.-P. Onnela, N. L. Keating und B. E. Landon. "Physician patient-sharing networks and the cost and intensity of care in US hospitals". In: *Medical care* 50.2 (2012), S. 152–160 (siehe S. 4).

[Bar07] M. J. Barber. "Modularity and community detection in bipartite networks". In: *Physical Review E* 76.6 (2007). URL: http://arxiv.org/abs/arXiv:0707.1616 (siehe S. 4, 9, 11).

[BB14] R. Busse und M. Blümel. "Germany: Health system review". In: *Health systems in transition* 16.2 (2014), S. 1–296, xxi (siehe S. 6).

[BE97] S. Borgatti und M. Everett. "Network Analysis of 2-Mode Data". In: (1997) (siehe S. 7).

[Bec16] S. J. Beckett. "Improved community detection in weighted bipartite networks". In: *Royal Society Open Science* 3.1 (2016). eprint: http://rsos.royalsocietypublishing.org/content/3/1/140536.full.pdf. URL: http://rsos.royalsocietypublishing.org/content/3/1/140536 (siehe S. 4, 5, 8, 9, 14, 16).

[Blo+08] V. D. Blondel, J.-L. Guillaume, R. Lambiotte und E. Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008 (siehe S. 4, 5, 9, 11, 13, 14).

[BM14] P. A. Bodoia M. Griffiths L. "Comparing Performance Across Paradigms of Community Detection in BipartiteNetworks". In: (2014). URL: http://snap.stanford.edu/class/cs224w-2014/projects2014/cs224w-40-final.pdf (siehe S. 4, 8).

[Bor09] S. Borgatti. "2-Mode Concepts in Social Network Analysis". In: Hrsg. von R. Meyers. 2009 (siehe S. 7).

[CNM04] A. Clauset, M. E. J. Newman und C. Moore. "Finding community structure in very large networks". In: *Physical Review E* 70 (2004), S. 066111. URL: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0408187 (siehe S. 14).

[FB07] S. Fortunato und M. Barthelemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1 (2007), S. 36–41 (siehe S. 11).

[FH16] S. Fortunato und D. Hric. "Community detection in networks: A user guide". In: *Physics Reports* 659 (2016). Community detection in networks: A user guide, S. 1 –44. URL: http://www.sciencedirect.com/science/article/pii/S0370157316302964 (siehe S. 4).

[For10] S. Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3 (2010), S. 75 –174. URL: http://www.sciencedirect.com/science/article/pii/S0370157309002841 (siehe S. 4, 5).

[GN02] M. Girvan und M. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Science* 99.12 (2002), S. 7821–7826 (siehe S. 3, 8, 10).

[Got+14] H. Gothe, P. Ihle, E. Swart und D. Matusiewicz. *Routinedaten im Gesundheitswesen - Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven*. Bern: Huber Hans, 2014 (siehe S. 7).

[GSPA07] R. Guimerà, M. Sales-Pardo und L. A. N. Amaral. "Module identification in bipartite and directed networks". In: *Phys. Rev. E* 76 (3 2007), S. 036102. URL: https://link.aps.org/doi/10.1103/PhysRevE.76.036102 (siehe S. 7, 8, 11).

[Lan+13] B. E. Landon, J.-P. Onnela, N. L. Keating, M. L. Barnett, S. Paul, A. J. O'Malley, T. Keegan und N. A. Christakis. "Using administrative data to identify naturally occurring networks of physicians". In: *Medical care* 51.8 (2013), S. 715–721 (siehe S. 4).

[LCJ14] D. B. Larremore, A. Clauset und A. Z. Jacobs. "Efficiently inferring community structure in bipartite networks." In: *CoRR* abs/1403.2933 (2014). URL: http://dblp.uni-trier.de/db/journals/corr/corr1403.html#LarremoreCJ14 (siehe S. 8).

[LM10] X. Liu und T. Murata. "An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks." In: *JACIII* 14.4 (2010), S. 408–415. URL: http://dblp.uni-trier.de/db/journals/jaciii/jaciii14.html#LiuM10 (siehe S. 8).

[New04] M. E. J. Newman. "Fast algorithm for detecting community structure in networks". In: *Phys. Rev. E* 69 (6 2004), S. 066133. URL: https://link.aps.org/doi/10.1103/PhysRevE.69.066133 (siehe S. 4).

[New10] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010 (siehe S. 3, 8).

[PCK16] P. Pesantez-Cabrera und A. Kalyanaraman. "Detecting Communities in Biological Bipartite Networks." In: *BCB*. ACM, 2016, S. 98–107. URL: http://dblp.uni-trier.de/db/conf/bcb/bcb2016.html#Pesantez-Cabrera16 (siehe S. 8).

[PL05] P. Pons und M. Latapy. "Computing communities in large networks using random walks". In: *International Symposium on Computer and Information Sciences*. Springer. 2005, S. 284–293 (siehe S. 14).

[SEC14] D. von Stillfried, M. Erhart und T. Czihal. "Ambulante Versorgung". In: *Medizinökonomie 1*. Hrsg. von C. Thilscher. 2014, S. 295–350 (siehe S. 6).

[Sti+17] D. von Stillfried, T. Ermakova, F. Ng und T. Czihal. "Virtuelle Behandlernetzwerke: Neue Ansätze zur Analyse und Veränderung räumlicher Versorgungsunterschiede". In: (Okt. 2017) (siehe S. 4).

[Uga+11] J. Ugander, B. Karrer, L. Backstrom und C. Marlow. "The Anatomy of the Facebook Social Graph". In: *CoRR* abs/1111.4503 (2011). arXiv: 1111.4503. URL: http://arxiv.org/abs/1111.4503 (siehe S. 3, 8).

[WT07] K. Wakita und T. Tsurumi. "Finding Community Structure in Mega-scale Social Networks". In: *CoRR* abs/cs/0702048 (2007). URL: http://dblp.uni-trier.de/db/journals/corr/corr0702.html#abs-cs-0702048 (siehe S. 14).

[ZA15] Z.-Y. Zhang und Y.-Y. Ahn. "Community detection in bipartite networks using weighted symmetric binary matrix factorization." In: *CoRR* abs/1502.04428 (2015). URL: http://dblp.uni-trier.de/db/journals/corr/corr1502.html#ZhangA15 (siehe S. 8).