TANJA ZAST

METHODEN DER ANALYSE VON SOZIALEN NETZWERKEN

universität ulm

METHODEN DER ANALYSE VON SOZIALEN NETZWERKEN

TANJA ZAST BACHELOR OF SCIENCE

Bachelor Thesis

Institute of Information Resource Management
Faculty of Engineering, Computer Science and Psychology
Ulm University

April 2022

Prof. Dr.-Ing. Dr. h.c. Stefan Wesner
Dr. Dipl.-Inf. Lutz Schubert

ZUSAMMENFASSUNG

Diese Arbeit handelt von sozialen Netzwerken, ihrer Generierung und anschließenden Analyse. Es werden Methoden vorgestellt, die zur Analyse benötigt werden und zudem die mathematischen Verteilungen dieser angewendeten Methoden betrachtet. Am Ende folgt ein Ausblick, über weitere Optimierungsmöglichkeiten der Generierung und Analyse von sozialen Netzwerken.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

Teil I

EINFÜHRUNG IN DIE THEORIE

Um das Thema zu verstehen und vor allem die spätere Interpretation, ist es nun von Bedeutsamkeit, eine Einführung in die Theorie zu ermöglichen.

1 EINLEITUNG

Der Begriff "Soziales Netzwerk" oder auf Englisch "Social Network" weckt seit vielen Jahrzehnten das Interesse zahlreicher Sozial- und Verhaltenswissenschaftler*innen [10]. Auch weckt es das Interesse von Unzähligen Unternehmen, um gezielter auf das Kundenverhalten einzugehen und dadurch den Gewinn zu maximieren [9]. Doch nicht zu vergessen sind es heutzutage letztendlich die Nutzer*innen der Social Media-Plattformen wie Twitter, Facebook und Instagram, welche dieser Begriff vor allem tangiert und die Liste könnte noch lange weitergeführt werden. Jedoch spezialisieren sich vor allem Sozial- und Verhaltenswissenschaftler*innen, ebenso Unternehmen, auf die Analyse sozialer Netzwerke [10][9]. Diese fokussieren sich weitestgehend auf Beziehungen zwischen sozialen Einheiten, sowie die Muster und Implikationen, welche diesen Beziehungen zugeschrieben werden [15]. Schnell kommen Fragen auf wie, was ist ein "Soziales Netzwerk" definiert. Oder wie eine solche Analyse aussehen kann. Was jede einzelne Methode zur Analyse auszeichnet und Welche davon als besonders vielversprechend gelten.

1.1 ZIELSETZUNG

Um eine Aussage darüber treffen zu können, welche Methoden zur Analyse geeignet sind und welche nicht, muss zunächst ein Verständnis für soziale Netzwerke und anschließende Analyse hergestellt werden. Diese Arbeit wird daher in zwei Bereiche unterteilt. Zum Einen beginnt sie mit der Einführung in die sozialen Netzwerke und die Einarbeitung in die verschiedenen Zentralitäten, die bei der Analyse verwendet werden. Diese geben einen guten Aufschluss darüber, wie die Einheiten miteinander verbunden sind beziehungsweise zusammenhängen. Ob es sich starke Verbindungen oder schwache handelt. Danach wird eine weitere Methode vorgestellt, welches es durch Zuordnung der Zentralitäten ermöglicht, die mathematische Gaußverteilung nachzustellen. Anhand dieser sind dann weitere Aussagen über den Graphen möglich. Nachdem ein Verständnis entwickelt wird, was verschiedene soziale Netzwerke auszeichnet und von random Graphen unterscheidet, wird anschließend im zweiten Teil dieser Arbeit ein Generator entwickelt, welcher Soziale Netzwerke so gut wie möglich nachstellt. Indem ein Generator entwickelt wird, der Testdaten beziehungsweise Test-Netzwerke erstellt, die zum Training oder für komperative Analysen genutzt werden können. Um aber bewerten können, ob dieses Netzwerk eine gute Simulation ist, werden die im ersten Teil der Arbeit vorgestellten Methoden angewendet. Ziel der Arbeit ist es daher, ein gutes Verständnis für die soziale Netzwerkanalyse zu bekommen und für beliebige Netzwerke, durch Anwendung der kennengelernten Methoden, gute Bewertungen oder Analysen durchzuführen. Diese Arbeit distanziert sich von dem Begriff "Social Networking", welcher bei Recherchen zahlreichst auftaucht, aber lediglich den Vorgang oder Zustand beschreibt, dass Menschen über soziale Netzwerke durch beispielsweise gemeinsame Interesse zueinanderfinden.

2 EINFÜHRUNG IN DIE SOZIALEN NETZWERKE

Um zu verstehen, warum Soziale Netzwerke analysiert werden, sollte zunächst die Frage geklärt werden, was ein Soziales Netzwerk ist. Hierfür existieren zwei Definitionen, eine gehört dem Bereich der Soziologie an und die andere dem Bereich des Internets.
In der Soziologie, ist ein soziales Netzwerk eine soziale Struktur, welche zwischen Akteuren besteht. Ein Akteur kann entweder von einer Einzelpersonen oder von Organisationen repräsentiert werden. Ein soziales Netzwerk zeigt die Art und Weise, wie Menschen und Organisationen durch soziale Vertrautheiten verbunden sind, die von zufälligen Bekanntschaften bis hin zu engen familiären Bindungen reichen [16]. Im Bereich des Internets ist der Begriff des Sozialen Netzwerks erst mit dem Web 2.0 entstanden. Der Begriff bezeichnet eine virtuelle Gemeinschaft. Diese wird überwiegend über die Internetplattform gepflegt und aufrechterhalten. Soziale Netzwerke variieren in ihren Funktionen. Beispiele hierfür sind themenorietierte Netzwerke, siehe Twitter, oder Netzwerke, die überwiegend der zwischenmenschlichen Kommunikation dienen, siehe Facebook [17]. Das heißt, die Soziologie bezeichnet ausschließlich die soziale Struktur, wohingegen im Internet die virtuelle Gemeinschaft bezeichnet wird.

2.1 ZIELE DER ANALYSE

Der Fokus der "Sozialen Netzwerkanalyse" liegt auf der Interpretation und Analyse sozialer Beziehungen. Genauer gesagt auf die Beziehungen zwischen zwei sozialen Einheiten. Forscher haben erkannt, dass die Netzwerkperspektive neue Erkenntnisse und Möglichkeiten zur Beantwortung sozial- und verhaltenswissenschaftlicher Standardforschungsfragen bietet. Dies ist möglich, da die "Soziale Netzwerkanalyse" das soziale Umfeld als Muster oder Regelmäßigkeiten in Beziehungen zwischen Einheiten ausdrücken, beziehungsweise darstellen kann. Das regelmäßige Muster in den Beziehungen kann auch als Struktur bezeichnet werden [18]. Die Analyse, welche im Folgenden behandelt werden misst diese Strukturen, wodurch genauere Aussagen oder auch Vermutungen über die Beziehungen getroffen werden können. Die Beziehungen in sozialen Netzwerken können unterschiedlicher Art sein, beispielsweise wirtschaftlich oder politisch, was nur zwei von vielen weiteren möglichen Beziehungstypen sind. Um die Muster oder Strukturen zu erkennen, erfordert es Methoden oder analytische Konzepte. In den letzten Jahrzehnten haben sich die Methoden zur Analyse von sozialen Netzwerken als großer Bestandteil der Fortschritte in der Sozialtheorie erwiesen. Die Analyse sozialer Netzwerke besteht aus einer Reihe von mathematischen und grafischen Verfahren beziehungsweise Techniken, welche Indizes zwischen Einheiten verwenden, um soziale Strukturen kompakt und systematisch darzustellen. Die Netzwerkanalyse verfolgt mehrere Ziele. Das erste Ziel ist die visuelle Darstellung von Beziehungen. Dies wird in Form eines Netzwerks oder Graphen abgebildet. Ein weiteres Ziel ist die Darstellung von Informationen. Dies soll es Benutzer*innen ermöglichen, die Beziehungen zwischen den Akteuren auf einen Blick zu erkennen. Zusätzlich verfolgt die Analyse das Ziel, grundlegende Eigenschaften von Beziehungen in einem Netzwerk

zu untersuchen. Dies sind Eigenschaften wie beispielsweise die Dichte und Zentralität. Ein weiteres Ziel besteht darin, Hypothesen über die Struktur der Verbindungen zwischen den Akteuren zu testen. Analysten sozialer Netzwerke können die Auswirkungen von Beziehungen auf die Einschränkung oder Verbesserung des individuellen Verhaltens oder der Netzwerkeffizienz untersuchen. Ein großer Vorteil von diesem Ansatz besteht darin, dass er sich auf die Beziehungen zwischen Akteuren konzentriert. Diese sind in ihren sozialen Kontext eingebettet. Soziale Netzwerkanalyse kann in vier Schritte unterteilt werden. Erstens in die Definition eines Netzwerks, Messung der Beziehungen, Darstellung der Beziehungen und schließlich die Analyse der Beziehungen [18]. Um diese Einteilung sinnvoll durchführen zu können, ist es von Vorteil, wenn die Netzwerke eine gewisse Grundstruktur aufweisen.

2.2 EINFÜHRUNG IN DIE GRUNDSTRUKTUR VON SOZIALEN NETZWERKEN

Ein Graph G, der aus disjunkten Mengen (V, E) besteht. Dabei bezeichnet V eine Menge von Knoten, und E stellt die sogenannten Kanten oder Bögen dar.
Wenn das Netz ungerichtet ist, d.h. für jede Verbindung, die von jedem Paar i nach j geht, gibt es eine Verbindung j nach i. Dies Verbindungen werden als Kanten bezeichnet. Andernfalls werden gerichtete Verbindungen als Bögen bezeichnet. Netzwerkkanten können auch Gewichte haben, die z.B. die Stärke der Interaktion zwischen zwei Knoten angeben. Soziale Netzwerke können entweder als Graphen oder Matrizen dargestellt werden. Eine Netzwerkmatrix ist eine quadratische Anordnung von Messungen, die das Vorhandensein oder Fehlen von Kommunikationsverbindungen zwischen Akteuren darstellen [3]. Das Vorhandensein wird mit einer "1" und das Nichtvorhandensein mit einer "0" beschrieben. Netzwerkmatrizen geben Verbindung zwischen den Knotenpunkten an. Da jede Adjazenzmatrix auch eine Netzwerkmatrix ist, ist in Zukunft von Adjazenzmatrizen die Rede.
Im Folgenden sind diese Matrizen zu betrachten:

Netzwerk 1:

$$\begin{pmatrix} & A & B & C & D & E \\ A & 0 & 0 & 0 & 1 & 1 \\ B & 1 & 0 & 1 & 1 & 1 \\ C & 0 & 1 & 0 & 1 & 0 \\ D & 1 & 1 & 0 & 0 & 1 \\ E & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Netzwerk 2:

$$\begin{pmatrix} & A & B & C & D & E \\ A & 0 & 1 & 0 & 1 & 1 \\ B & 0 & 0 & 1 & 0 & 1 \\ C & 1 & 1 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 1 \\ E & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Die erste Spalte und die erste Zeile der beiden Matrizen, stellt die Knoten innerhalb des Netzwerks dar. In sozialen Netzwerken ist es eher untypisch, dass Knoten auf sich selbst abbilden. Das würde beispielsweise heißen, dass eine Person eine Verbindung zu sich selbst aufweist, sich selbst folgt, oder die eigenen Beiträge liked, was üblicherweise nicht der Fall ist. Daher stehen in den beiden oberen Matrizen in den Diagonalen immer die Ziffer 0. Das heißt,

es sind keine Kanten vorhanden vom Knoten zu sich selbst [18].
Jedoch war die Rede davon, dass soziale Netzwerke nicht nur in Form von Matrizen dargestellt werden können, sondern auch also Graphen abgebildet werden. Die Matrizen oben bieten sich dafür idealerweise an. Die Graphen würde in diesem Fall wie folgt aussehen:

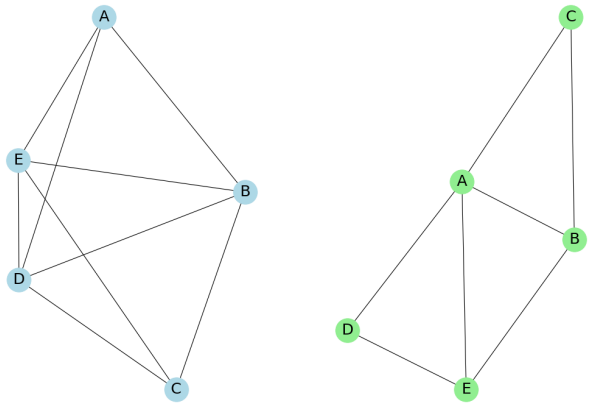


Abbildung 2.1: Links ist Netzwerk1 als Graph dargestellt und rechts Netzwerk2

Daher können für jegliche Netzwerkanalysen beide Varianten verwendet werden. Jedoch werden in dieser Arbeit überwiegend Graphen zur Veranschaulichung und Matrizen für jegliche Berechnungen verwendet, da es leichter ist auf den Datentyp einer Matrix zuzugreifen.

Ein soziales Netzwerk ist eine soziale Struktur, die zwischen Akteuren - Einzelpersonen oder Organisationen - besteht. Es zeigt die Art und Weise, wie Menschen und Organisationen durch verschiedene soziale Vertrautheiten verbunden sind, die von zufälligen Bekanntschaften bis hin zu engen familiären Bindungen reichen. Soziale Netzwerke bestehen aus Knotenpunkten und Verbindungen, deren Wechselwirkung nicht linear ist. Die Person oder Organisation, die am Netzwerk teilnimmt, wird als Knoten bezeichnet. Bindungen sind die verschiedenen Arten von Verbindungen zwischen diesen Knotenpunkten. Bindungen werden nach ihrer Stärke bewertet. Lockere Verbindungen, wie bloße Bekanntschaften, werden als schwache Verbindungen bezeichnet. Starke Verbindungen, wie z. B. Familien oder Cliquen, werden als starke Bindungen bezeichnet [16]. Beispiele für soziale Netzwerke sind unsere Gesellschaft, das Internet, unser Gehirn und zelluläre Interaktionen. Doch welche grundsätzlichen Eigenschaften muss ein Netzwerk erfüllen, um als ”soziales Netzwerk” bezeichnet werden zu dürfen? Sozialwissenschaftler*innen haben drei Arten von Netzwerken untersucht: egozentrische, soziozentrische und systemoffene Netzwerke. Egozentrische Netze sind Netze, die mit einem einzigen Knoten oder einer einzigen Person verbunden sind [7]. Um als Netze zu gelten, müssen diese Verbindungen nicht nur Listen von Personen oder Organisationen sein, sondern es müssen auch Informationen über die Verbindungen zwischen diesen Personen oder Organisationen enthalten sein. Im allgemeinen Sprachgebrauch, insbesondere wenn von sozialer Unterstützung

die Rede ist, wird jede Liste als Netzwerk betrachtet. Eine Person, die eine große Anzahl guter Freunde hat, auf die sie auf die sie sich verlassen kann, wird ein großes "Netzwerk"genannt. Soziozentrische Netzwerke sind, wie Russell Bernard Bedeutung (persönliche Kommunikation), Netzwerke in einer Box. Netze mit offenen Systemen sind Netze, bei denen die Grenzen nicht klar sind, sie liegen nicht in einer Box - zum Beispiel die Verbindungen zwischen Unternehmen, oder die Kette an Auswirkungen, die eine Entscheidung oder deine Erneuerung in beispielsweise technischen Prozessen nachzieht. In gewisser Weise sind dies die interessantesten Netzwerke. Sie sind auch am schwierigsten zu untersuchen [4].

3 KERNFAKTOREN EINER SOZIALEN NETZWERKANALYSE

In komplexen Netzwerken können einige Knoten als wichtiger angesehen werden als andere. In einem sozialen Netzwerken zeichnen sich manche Knoten durch vergleichsweise mehr Verbindungen als andere Knoten aus. Auf das Beispiel *Instagram* bezogen, können solche Knoten Informationen leichter verbreiten, als gewöhnliche Personen, sogenannte Influencer. Daher können diese Knotenpunkte als zentral interpretiert werden. Die Interpretation der Zentralität ist jedoch nicht eindeutig [2]. Zum Beispiel im Linienverkehr, gilt eine Linie als zentral, wenn sie von großen Menschenmengen genutzt wird und stärker frequentiert wird als andere Linien. Die Definition der Zentralität ist also nicht allgemein und hängt von der der Anwendung ab. Da es keine allgemeine Definition von Zentralität gibt, wurden mehrere Maße entwickelt, die jeweils spezifische Konzepte berücksichtigen. Die Zentralität ist eine Schlüsseleigenschaft komplexer Netzwerke. Sie kann unter anderem das Verhalten dynamischer Prozesse und beispielsweise epidemische Ausbreitung erklären, modellieren und abschätzen, jedoch nicht beschreiben. [6]. Zudem kann die Zentralität Informationen über die Organisation komplexer Systeme, wie unser Gehirn, und unsere Gesellschaft liefern. Es gibt viele Metriken zur Quantifizierung der Knotenzentralität in Netzwerken [12]. Nun folgt ein Überblick, über die wichtigsten Zentralitätsmaße und die Hauptmerkmale dieser.

3.1 ZENTRALITÄTEN

Die Gradzentralität ist die am einfachsten zu berechnende Zentralität. Sie ist definiert durch die Anzahl der Verbindungen, die mit jedem Knoten *direkt* verbunden sind. Mit der Adjazenzmatrix wird der Grad der Zentralität berechnet, indem die Summe der Elemente der betroffenen Zeile i berechnet wird. Mathematisch formuliert, wird folgende Formel verwendet:

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{3.1}$$

Wobei A die Adjazenzmatrix beschreibt, N die Anzahl an Knoten darstellt und i, j die Knoten.
Da es sich bei der Gradzentralität um die einfachste Zentralität handelt, wird meist davon ausgegangen, dass Knoten mit vielen Verbindungen, daher mit einer hohen Zentralität, sich im Zentrum eines Netzwerkers befinden. Dies hat jedoch einige Nachteile, denn Knoten mit der höchsten Grad-Zentralität können sich auch am Rand des Netzes befinden, sich daher nicht im Zentrum befinden, was dazu führt, dass die Gradzentralität nicht als lokales Maß betrachtet wird. Zudem sollte hervorgehoben werden, dass bei der Gradzentralität nur ein- beziehungsweise ausgehende Kanten gezählt werden. Die sagt zwar aus, dass ein solcher Knoten, auf das soziale Netzwerk bezogen, eine beliebte oder sehr bekannte Person ist, doch ist keine Aussage über die Macht oder den Einfluss der Person ermöglicht. Als extremes Gegenbeispiel, warum die Gradzentralität nicht immer optimal zur Netzwerkanalyse ist, diene ein Netzwerk

mit einer großen, dichten Gruppen von Knoten. Diese machen den größten Teil des Graphen aus, was auch manchmal als Kern des Netzes bezeichnet wird. Jedoch kann (visuell betrachtet) weit außerhalb des Kerns, entlang einer Kette von Knoten mit niedrigem Grad, ein Knoten liegen, der mit einer großen Anzahl von Knoten verbunden ist verbunden ist. Ein solcher Knoten hätte einen hohen Grad an Zentralität, obwohl er weit vom Kern des Netzes und den meisten Knoten entfernt ist, in visueller Hinsicht [6]. Um solche Faktoren mit berücksichtigen zu können, wird ein weiterer Faktor mit in die Berechnung integriert, nämlich die Weglänge.

Diese spielt eine wichtige Rolle bei der "Nähe-Zentralität". Denn die Knotenzentralität kann auch anhand der kürzesten Wege definiert werden. Der Abstand zwischen zwei Knoten i und j ist gegeben durch die Anzahl der Kanten, welcher sie verbindet. Ein zentraler und daher wichtiger Knoten liegt, bezogen auf den Abstand, nahe an allen anderen Knoten des Netzes. Dieser Gedanke ist im Maß der sogenannten "Nähe-Zentralität" oder "Closeness-Centrality" enthalten. Diese wird durch den durchschnittlichen Abstand eines jeden Knotens zu allen anderen Knoten definiert. Mathematisch wird die Formel wie folgt beschrieben:

$$C_i = \frac{N}{\sum_{j=1, j\neq i}^{N} d_{ij}} \tag{3.2}$$

Dabei ist mit d_{ij} der kürzeste Weg zwischen i und j gemeint und mit N erneut die Anzahl an Knoten im Netzwerk [6]. Die Nähe-Zentralität ist vor allem dann sehr geeignet, wenn Prozesse über kurze Wege charakterisiert werden wollen. Beispielsweise kann der hierarchischen Aufbau eines Unternehmens in einem sternförmigen Graphen dargestellt werden. In der Mitte des Graphen befindet sich der Vorstand, der in engem Kontakt mit den jeweiligen Abteilungsleitern steht. Die Abteilungsleiter sind, neben dem Vorstand, wiederum in sehr nahem Kontakt mit ihren jeweiligen Mitarbeitern ihrer Abteilung. Wenn nun ausschließlich anhand der Grad-Zentralität argumentieren wird, sind die Abteilungsleiter die wichtigsten Knoten im Graphen. Jedoch haben diese nicht die niedrigste Nähe-Zentralität, denn der Vorstand hat, da sich dieser Knoten in der Mitte des Graphen befindet, zu allen anderen Knoten entweder einen oder zwei Kanten Abstand. Die einzelnen Abteilungsleiter haben aber im worst-case zu anderen Angestellten aus anderen Abteilungen zwei bis drei Kanten Abstand. Dementsprechend ist es wichtig, auch die Nähe-Zentralität zu betrachten, denn diese ist von hoher Bedeutung. Tatsächlich weisen die meisten komplexen Netze eine geringe durchschnittliche Länge des kürzesten Weges auf. Dies ist dadurch zu begründen, da die typische Entfernung mit dem Logarithmus der Anzahl der Knoten zunimmt. Daher liegt das Verhältnis zwischen dem größten und dem kleinsten Abstand in der Größenordnung $log(N)$, da der minimale Abstand gleich eins ist. In den meisten real existierenden Netzwerken beträgt dieses Verhältnis etwa sechs oder weniger. Es kann also mehrere Knoten mit der gleichen Zentralität haben, obwohl sie bei der Informationsverbreitung unterschiedliche Rollen spielen können. Daher ist die Nähe-Zentralität besser geeignet für räumliche Netze, bei denen die Abstände zwischen den Knoten größer ist als in zufälligen Netzen mit der gleichen Anzahl von Knoten und Verbindungen [6].

Die Bewteenesss-Zentraität hingegen misst, wie wichtig ein Knoten für die kürzesten Pfade durch das Netz ist. Um diese Zentralität für einen Knoten N zu berechnen, wird in dieser Methode eine Gruppe Knoten ausgewählt und alle kürzesten Wege zwischen diesen Knoten gefunden. Dann wird der Anteil dieser kürzesten Wege berechnet, die den Knoten N einschließen. Wenn es beispielsweise sieben kürzeste Wege zwischen einem Knotenpaar gibt und

fünf davon durch den Knoten N führen, dann wäre der Anteil $5/7 = 0.714$. Dieser Vorgang wird für jedes Knotenpaar im Netz wiederholt. Anschließend werden die berechneten Bruchteile addiert, wodurch die Betweeness-Zentralität des Knotens N generiert wird. Mathematisch formuliert sieht die Formel dann wie folgt aus:

$$B_i = \sum_{(a-b)} \frac{\eta(a,i,b)}{\eta(a,b)} \tag{3.3}$$

Hierbei bezeichnet $\eta(a,i,b)$ die Anzahl der kürzesten Wege zwischen den Knoten a und b die durch den Knoten i führen. Zudem stellt $\eta(a,b)$ die Gesamtzahl der kürzesten Wege zwischen a und b dar. Diese Zentralität, basierend auf dem " random walk"-Algorithmus, ist gegeben durch die erwartete Anzahl der Besuche jedes Knotens i während einer zufälligen Schrittfolge durch den Graphen:

$$B_i = \sum_{a=b}^{N} \sum_{b=1}^{N} w(a,i,b) \tag{3.4}$$

dabei ist $w(a,i,b)$, wie oben bereits beschrieben für $\eta(a,i,b)$, die Anzahl der kürzesten Wege zwischen den Knoten a und b die durch den Knoten i führen. Die Lösung wird nur angenähert. Die Betweeness-Zentralität ist eines der am häufigsten verwendeten Zentralitätsmaße. Sie gibt an, wie wichtig ein Knoten für den Informationsfluss von einem Knoten des Netzes zu einem anderen ist. In gerichteten Netzwerken kann Betweenness mehrere Bedeutungen haben [6]. Einem Nutzer mit hoher Betweeness-Zentralität folgen möglicherweise viele andere Nutzer, die jedoch nicht denselben Personen folgen wie der Nutzer selbst. Dies würde darauf hindeuten, dass der Nutzer viele Anhänger oder Follower hat. Es kann aber auch sein, dass der Nutzer weniger Follower hat, diese aber dafür mit vielen Konten verbindet, die ansonsten weit entfernt sind. Dies würde darauf hindeuten, dass der Nutzer ein Anhänger von vielen Personen ist, beziehungsweise vielen Personen folgt. Daher ist es enorm wichtig die Richtung der Kanten eines Knotens zu kennen, um die Bedeutung der Zentralität zu verstehen.

Die Eigenvektor- oder Eigenwert-Zentralität misst die Bedeutung eines Knotens, wobei die Bedeutung seiner Nachbarn berücksichtigt wird. Daher wird sie manchmal verwendet, um den Einfluss eines Knotens im Netzwerk zu messen. Er wird durch eine Matrixberechnung ermittelt, um den so genannten "*Haupteigenvektor*" anhand der Adjazenzmatrix zu bestimmen. Mathematisch betrachtet ist die Eigenvektor-Zentralität die komplizierteste, der in dieser Arbeit betrachteten Zentralitäten.
Wird nun von die Tatsachte betrachtet, dass ein Akteur zentraler ist, wenn er in Beziehung zu Akteuren steht, die selbst zentral sind. So kann also argumentiert werden, dass die Zentralität eines Knotens nicht nur von der der Anzahl seiner Nachbarknoten abhängt, sondern auch von deren Zentralitätswert. Beispielsweise definiert Bonacich (1972) die Zentralität $c(v_i)$ eines Knotens v_i als positives Vielfaches der Summe der benachbarten Zentralitäten. Als Formel mathematisch dargestellt sieht dies folgendermaßen aus:

$$\lambda c(v_i) = \frac{1}{\lambda} \sum_{j=1}^{N} a_{ij} c(v_j) \forall i \tag{3.5}$$

oder umgeschrieben:

$$c(v_i) = \sum_{j=1}^{N} a_{ij} c(v_j) \forall i \tag{3.6}$$

Hierbei repräsentiert $a_{i,j}$ die Werte der Adjazenzmatrix A und λ einen konstanten Faktor.
In Matrixschreibweise mit $c = (c(v_1), ..., c(v_n))$ bedeutet dies auch:

$$Ac = \lambda c \tag{3.7}$$

Diese Art von Gleichung wird durch die Eigenwerte und Eigenvektoren von A gelöst. Aus der gesamten Menge an verschiedenen Eigenvektoren, scheint nur einer eine geeignete Lösung zu sein. Dieser Eigenvektor kann dann direkt als Zentralitätsmaß dienen. Da A die Adjazenzmatrix eines ungerichteten (zusammenhängenden) Graphen ist, ist A nicht negativ und aufgrund des Satzes von Perron-Frobenius, gibt es einen Eigenvektor des maximalen Eigenwerts mit nur nicht negativen, also positiven, Einträgen [13].

3.2 CLIQUEN UND BRÜCKEN

Eine Clique ist laut Definition ein Teilgraph, aus mindestens drei Knoten bestehend, die zudem alle benachbart zueinander sind, auch streng bezeichnet als eine zusammenhängende Untergruppe. Eine Clique kann als Ansammlung von Akteuren gesehen werden, die sich gegenseitig *wählen*, jedoch *wählt* kein anderer Akteur dieser Gruppe und wird auch nicht von allen anderen Akteuren *gewählt*. Es ist zu beachten, dass sich Cliquen in einem Graphen auch überlappen können, also derselbe Satz von Knoten zu mehr als nur einer Clique gehören kann. Jedoch kann eine vollständige Clique nicht in einer anderen Clique enthalten sein, denn sonst wäre die kleinere Clique nicht maximal. Die Cliquendefinition ist vor allem sehr nützlich für die Untersuchung der Eigenschaften einer Untergruppe beziehungsweise eines Subgraphen [18]. Was genau damit gemeint ist, ist in folgendem Plot zu sehen:

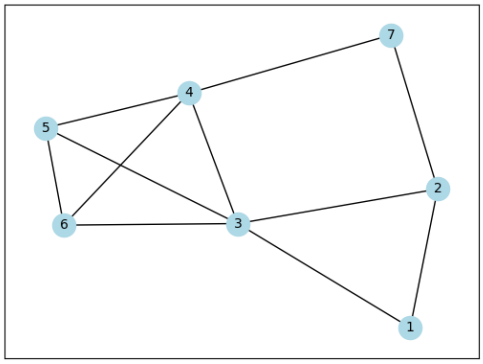


Abbildung 3.1: Graph mit den Cliquen (1, 2, 3) und (3, 4, 5, 6)

Wichtig ist hierbei, dass es sich bei (2, 3, 4, 7) um keine Clique handelt, da keine Verbindung zwischen den Knoten **4 und 2** und ebenso keine Verbindung zwischen den Knoten **3 und 7** besteht. Neben den Cliquen sind auch Brücken eine wichtige Diskussions- und Analysierungsgrundlage für Graphen beziehungsweise in unserem Fall für *soziale Netzwerke*. Wenn von Brücken (bzw. englisch Bridge) die Rede ist, sind Verbindungen zwischen zwei Knoten gemeint. Jedoch handelt es sich um die einzige Verbindung zwischen diesen Knoten und deren Kontakten [14]. Ein Beispiel für Brücken im Graphen liefert folgender Plot:

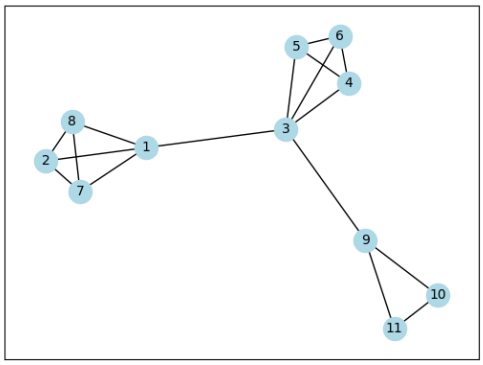


Abbildung 3.2: Graph mit den Brücken (1, 3) und (3, 9)

Hierbei ist gut zu erkennen, dass die drei Subgraphen durch *Brücken* miteinander verbunden sind. Neben den Brücken sind im Graphen 3.2 die Cliquen (1,2,7,8), (3,4,6,5) und (9,10,11) enthalten. Zu beachten ist, dass es sich bei beispielsweise (1,7,8) oder (4,5,6) um keine Cliquen handelt. Wenn nun diese Brücken und Cliquen im Zusammenhang zu Zentralitäten betrachtet werden und die oben aufgeführten Formeln der Zentralitäten auf den Graphen 3.2 angewendet werden erhält man die unten stehende Tabelle. Eine Berechnung für 3.1 ist nicht nötig, da in 3.2 ebenfalls Cliquen und zusätzlich Brücken enthalten sind:

Direkt fällt auf, dass die Werte spaltenweise sehr ähnlich zueinander sind. Bei der *Gradzentralität* sind die Knoten **3** und **1** mit einem Wert von **0.5** und **0.4** am höchsten. Interessant, denn dabei handelt es sich um die Knoten, die unsere *Brücke* bilden. Bei den Knoten **3** und **1** fällt auch weiter auf, dass diese Knoten bei der *Nähe-*, *Zwischen-* und *Eigenvektor-Zentralität* ebenfalls am höchsten sind. Das heißt, die Vermutung liegt nahe, dass Graphen, die Cliquen enthalten haben, relativ ähnliche *Zentralitätswerte* aufweisen beziehungsweise die Varianzen geringer sind. Aber vor allem erwähnenswert ist, dass in der Tabelle 3.1 lediglich bei den Knoten, welche die *Brücke* bilden, Werte ungleich *Null* auffindbar sind. Dies sollte für den weiteren Teil der Arbeit in Erinnerung bleiben.

3.3 EIN TYPISCHES SOZIALES NETZWERK

Nachdem nun alle Zentralitäten und deren Berechnungen bekannt sind, ist es an der Zeit ein Musterbeispiel für ein soziales Netzwerk zu betrachten. Google Maps ist beispielsweise

Tabelle 3.1: Werte Abbildung 3.2

Node	Grad-Zentr.	Nähe-Zentr.	Betweeness-Zentr.	Eigen.-Zentr.
3	0.5	0.666667	0.733333	0.470733
1	0.4	0.555556	0.466667	0.387253
4	0.3	0.454545	0	0.340195
6	0.3	0.454545	0	0.340195
5	0.3	0.454545	0	0.340195
2	0.3	0.4	0	0.279871
7	0.3	0.4	0	0.279871
8	0.3	0.4	0	0.279871
9	0.3	0.5	0.355556	0.184986
10	0.2	0.357143	0	0.0776041
11	0.2	0.357143	0	0.0776041

ein Netzwerk, bei dem die Knoten die *Orte* und die Kanten die *Straßen* sein können. Das bekannteste Netzwerk ist natürlich Facebook. Bei dieser sozialen Plattform ist die geeignetste Darstellung ein *ungerichteter* Graph. Bei Instagram hingegen, ein *gerichtet* Graph. Denn hier gibt es neben Leuten, denen wir folgen, unsere eigenen Follower [8]. Die Knoten sind die *Nutzer* und die *Kanten* sind die Verbindungen zwischen ihnen. Beachten Sie, dass sowohl *Knoten* als auch *Kanten* Attribute haben können. Knotenattribute in Facebook können zum Beispiel *Geschlecht*, *Ort*, *Alter* usw. sein, und Kantenattribute können *Datum der letzten Unterhaltung zwischen zwei Knoten*, *Anzahl der Likes*, *Datum der Verbindung* usw. sein [11]. Im folgenden wird ein, auf den ersten Blick erscheinendes, typisches soziales Netzwerk betrachtet. Es muss jedoch stets klar sein, dass es sich hierbei um einen Datensatz von einem fiktives Fantasy Drama handelt [11]:

Für diesen Plot wurde die "*NetworkX*" Python-Bibliothek auf "Game of Thrones"-Daten (GOT) angewendet. Das Netzwerk besteht aus 796 Knoten und 2823 Kanten. Insgesamt daher aus 796 Charakteren aus GOT.
In dieser SNA tauchen auch bisher unbekannte Messungen auf, die aber im Interpretations-Teil dieser Arbeit ebenfalls aufgegriffen werden. Beispielsweise beträgt der "Durchmesser" des GOT Graphen 9. Die heißt, wenn die kürzeste Pfadlänge von jedem Knoten zu allen anderen Knoten berechnet ist, ist der Durchmesser die längste aller berechneten Pfadlängen. Die durchschnittlich kürzeste Pfadlänge beträgt 3.41. Diese wird aber zu einem späteren Zeitpunkt analysiert. Im Graphen ist gut zu erkennen, welche Knoten eine zentrale Rolle in diesem Graphen spielen. Hierfür wird mit der Knoten-Größe variiert. Große Knoten implizieren, dass es sich um einen wichtigen Knoten für diesen Teilgraphen handelt und kleine, dass es sich um weniger relevante Knoten handelt [11]. Wenn diese in der Abbildung 3.3 gesucht werden, wird ersichtlich, dass es sich hierbei um die Knoten handelt, die mit den meisten Kanten verbunden sind. Oftmals ist bei den Graphen nicht eindeutig zu erkennen, ob es sich hierbei um Kanten handelt, die zum Knoten führen und sozusagen eine eingehende Kante darstellen, oder diese nur am Knoten vorbei verlaufen. Deshalb ist es wichtig, die Werte aus der Tabelle 3.2 zu analysieren. Hingegen fällt bei der Spalten "*Charakter*" auf, dass "*Tyrion – Lannister*" in allen Spalten aufgeführt wird. Das heißt, dass dieser Knoten im Graphen sowohl zentral liegen muss, zudem kurze

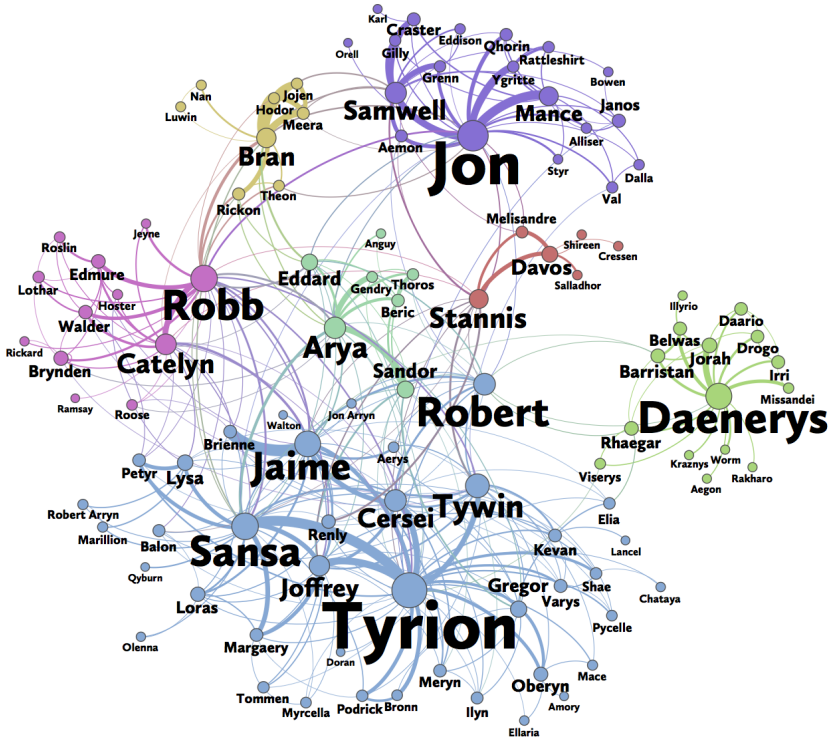


Abbildung 3.3: Game of Thrones social Network,
Quelle: https://predictivehacks.com/social-network-analysis-of-game-of-thrones/,
Stand: 28.03.2022

Tabelle 3.2: Werte GOT Graph

Charakter	Grad-Zentr.	Charakter	Nähe-Zentr.	Charakter	Betweeness-Zentr.
Tyrion Lannister	0.1535	Tyrion Lannister	0.4763	Jon Snow	0.1921
Jon Snow	0.1434	Robert Baratheon	0.4593	Tyrion Lannister	0.1622
Jaime Lannister	0.1270	Eddard Stark	0.4558	Daenerys Targaryen	0.1184
Cersei Lannister	0.1220	Cersei Lannister	0.4545	Theon Greyjoy	0.1113
Stannis Baratheon	0.1119	Jaime Lannister	0.4520	Stannis Baratheon	0.1101

Abstände zu den anderen Knoten nachweisen und über diesen Knoten verlaufen zudem die häufigsten kürzesten Wege. Bei dem Graphen 3.3 fällt auf, dass der Knoten, beziehungsweise Charakter, *Tyrion* heraus sticht. Er liegt zwar nicht komplett mittig im Graphen aber ist von den meisten Knoten und Kanten umgeben. Da drei der fünf wichtigsten Konten in der Spalte $Grad - Zentr.$ den gleichen zweiten Namen tragen, liegt die Vermutung nahe, dass es sich hier um Knoten handelt, die auch sehr nah beieinander sein müssten. Beim Betrachten des Graphen bestätigt sich diese Vermutung erneut, denn alle drei Knoten befinden sich im blauen Teilgraphen. Zudem haben Recherchen ergeben, dass es sich bei dem Namen "*Lannister*" um

ein Adelshaus in der US-amerikanischen Fantasy-Fernsehserie "Game of Thrones" handelt. Zudem fällt sofort auf, dass drei der fünf Charaktere in der Spalte *Nähe-Zentr.* bereits die selben sind, wie die wichtigsten Charaktere bezüglich der $Grad - Zentr$. Wieder bedeutet das, dass diese Charaktere sowohl zentral im Graphen liegen müssen und zudem die kürzesten Wege zu anderen Knoten besitzen. Die Betrachtung von 3.3 bestätigt dies sofort. Zudem weist der Graph auch einige cliquen auf, die relevanteste und vor allem größte Clique befindet sich im blauen, grünen, ein Knoten im roten und zwei Knoten im pinken Teilgraphen. Dies kann behauptet werden, weil aus dem Kapitel über Brücken und Cliquen bekannt ist, dass die Knoten mit den höchsten Zentralitäten in Cliquen enthalten sein müssten und es sich vor allem bei den Knoten mit hohen *Zwischen-* beziehungsweise *Betweeness-Zentralitäten* um um Brücken handelt. Jedoch wird die Analyse dieses sozialen Netzwerks nicht weiterführen, sondern weiter auf die Analyse des künstlich erstellen sozialen Netzwerks fokussieren. Auch der Frage welcher mathematische bzw. stochastische Verteilung die Zentralitäten entsprechen und warum eine solche Untersuchung sinnvoll ist, wird zu einem späteren Zeitpunkt nachgegangen.

3.4 KURZES RECAP

Nun sind die wichtigsten Eigenschaften der, in dieser Arbeit betrachteten und verwendeten, Zentralitäten bekannt und eingeführt. Manche Zentralitäten wurden oberflächlicher erklärt als andere, was die simple Begründung hat, dass sie weniger relevant für die Untersuchung der sozialen Netzwerke sind. Schließlich wurde ein Beispiel für ein soziales Netzwerk betrachtet und dieses oberflächlich analysiert.

Teil II

DER PRAKTISCHE TEIL

Nun folgt der Teil der Arbeit, in dem selbst generierte soziale Netzwerke untersucht werden. Handelt es sich bei den generierten Netzwerken tatsächlich um soziale Netzwerke, erfüllen sie alle Ansprüche bezüglich der Zentralitäten und sonstigen Eigenschaften von sozialen Netzwerken? Dies sind einige Fragen, die in diesem zweiten Teil der Arbeit beantwortet werden sollen.

4 DER GRAPHEN GENERATOR

Da nun die Theorie hinter sozialen Netzwerken und der Analyse dieser erarbeitet ist, beschäftigt sich diese Arbeit im weiteren damit, wie typische soziale Netzwerke generiert werden können. Zunächst bietet es sich oftmals an, da Facebook und Instagram der Informationspflicht unterliegen, die eigenen social Media Daten anzufordern. Meist spiegelt dieser Datensatz gelikete und kommentierte Posts der Nutzer*innen wieder, oder verfasste Nachrichten und gesuchte Inhalte. Bei den ersten Visualisierungsversuchen wird bereits klar, dass diese Daten für eine wissenschaftliche Arbeit nicht brauchbar sind, da es sich bei den erstellten Plots und Ergebnissen nicht um *typische soziale Netzwerke* handelt. Vielmehr bestehen diese meist aus einem Kernknoten, also einem sogenannten sternförmigen Graphen. Aber auch die Abbildung 4.1 ist typisch für die Visualisierung der eigenen Daten. Diese besteht aus unzähligen einzelnen Teilgraphen, welche lediglich eine weitere Verbindung aufweisen. Auch sind keine Cliquen oder Bridgen (Brücken) in solchen Graphen zu finden, was ebenfalls dafür spricht, dass es sich um kein *typisches soziales Netzwerk* handelt.

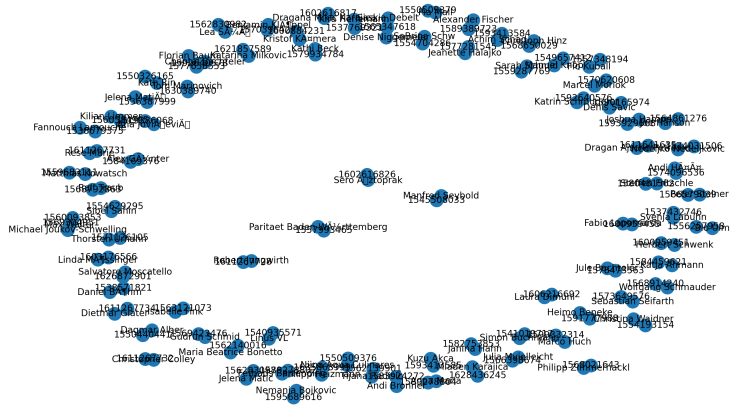


Abbildung 4.1: Erste Versuche eines Sozialen Netzwerks, selbst erstellt

Eine weitere Schwierigkeit ist bei diesen Graphen die Interpretation der Kanten, denn diese sind teilweise nicht eindeutig. Facebook gibt lediglich bestimmte IDs bekannt, doch welche exakte Bedeutung diese haben bleibt unklar.

4.1 GENERIERUNG EINES SOZIALEN NETZWERKS

Bei einer endlichen Anzahl von Knoten n gibt es auch eine endliche Anzahl von Graphen, die aus diesen Knoten erzeugt werden können. Hierbei wächst die Anzahl der Graphen mit n Knoten exponentiell. Ein Zufallsgraph ist nur einer dieser Graphen, der durch einen Zufallsprozess erzeugt werden kann. Wenn von *Zufallsgraphen* die Rede ist, wird in den meisten Fällen das *Erdős-Rényi-Modell* als Graphengenerator verwendet (benannt nach den Mathematikern Paul Erdős und Alfréd Rényi). Eine wichtige Eigenschaft von, auf diese Weise erzeugten Zufallsgraphen ist, dass alle Konstellationsmöglichkeiten des Graphen gleichverteilt erzeugt werden [1]. Neben dem Erdős-Rényi-Modell, gibt es noch viele weitere Methoden zur random Netzwerkmodellierung [1].

- die " dense_gnm_random_graph" liefert einen $G_{n,m}$-Zufallsgraphen. Bei dem $G_{n,m}$-Modell wird ein Graph gleichmäßig zufällig aus der Menge aller Graphen mit n Knoten und m Kanten ausgewählt.
- bei der "Newman–Watts–Strogatz small-world graph"-Modellierung wird zunächst ein Ring mit n Knoten erzeugt. Dann wird jeder Knoten im Ring mit seinen k nächsten Nachbarn verbunden (oder $k-1$ Nachbarn, wenn k ungerade ist). Anschließend wird für jede Kante (u, v) im zugrundeliegenden " n-Ring mit k nächsten Nachbarn", mit der Wahrscheinlichkeit p, eine neue Kante (u, w) mit einem zufällig ausgewählten bestehenden Knoten w hinzugefügt. Im Gegensatz zu "watts_strogatz_graph()" werden bei dieser Methode keine Kanten entfernt
- Die "random_regular_graph "-Modellierung gibt einen zufälligen d-regulären Graphen mit n Knoten zurück. Der resultierende Graph hat keine Selbstschleifen oder parallele Kanten
- Die "barabasi_albert_graph"-Modellierung hingegen liefert einen Zufallsgraphen nach dem Barabási-Albert-Präferenzmodell. Ein Graph mit n Knoten wird durch Anhängen neuer Knoten mit jeweils m Kanten erzeugt, die bevorzugt an bestehende Knoten mit hohem Grad angehängt werden.
- Die "powerlaw_cluster_graph"-Modellierung ist im wesentlichen das Barabási-Albert (BA)-Wachstumsmodell mit dem zusätzlichen Schritt, dass für jede zufällige Kante die Chance besteht, dass ebenfalls eine Kante zu einem seiner Nachbarn besteht (und damit ein Dreieck entsteht) [1].

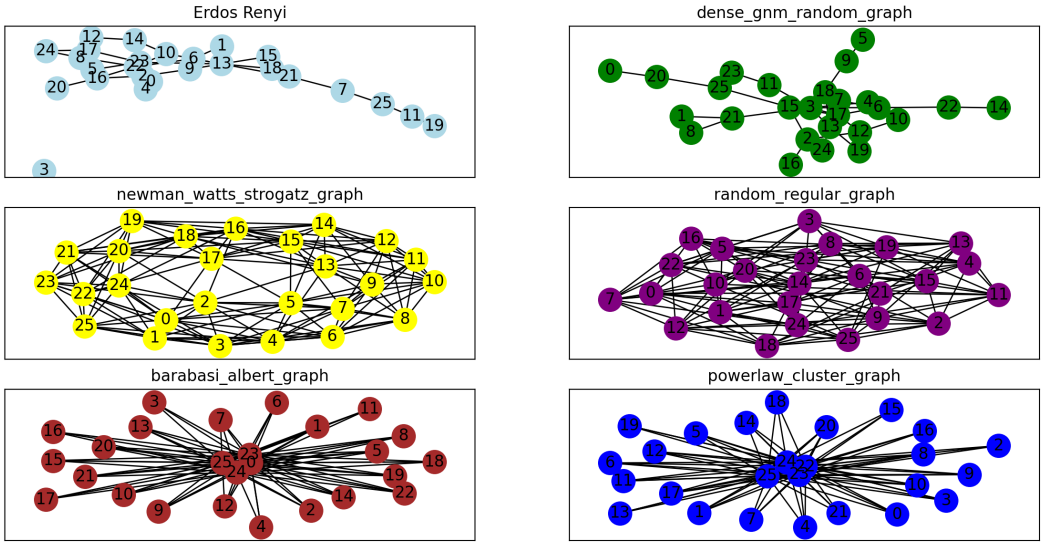


Abbildung 4.2: zufällig erstellte Graphen mit 25 Knoten nach den jeweiligen Methoden

Bei den Graphen 4.2 wurde lediglich eine visuelle Interpretation durchgeführt und nicht den Graphen mit den jeweiligen Zentralitäten analysiert. Auf den ersten Blick erkennen wir, dass bei allen sechs Modellen Unstimmigkeiten zu *sozialen Netzwerken* auftreten. Beispielsweise bei dem *Barabasi Albert Graph* unten links und dem *Powerlaw cluster graph* unten rechts sind einzelne zentrale Knoten zu erkennen. Diese zentrale Knoten befinden sich in der Mitte des Plots und sind von vielen weiteren Knoten umgeben, die alle wiederum mit diesen zentralen Knoten verbunden sind. Auch der *newman watts strogatz graph* und der *random regular graph* entsprechen nicht den erwünschten *sozialen Netzwerken*. Beide mittigen Plots sind ringförmig angeordnet und es scheint, als sei jeder Knoten mit jedem weiteren Knoten verbunden, was eine untypische Eigenschaft ist. Nun bleiben noch die beiden oberen Plots des *Erdos Renyi-Graphen* und *dense gnm random graph*, welche ebenfalls nicht unseren Erwartungen entsprechen. Der Plot des *dense gnm radom graph* weist zwar einzelne Äste auf, die aus der Mitte des Plots verlaufen, doch generell wenige Cliquen enthalten und keine Cluster aufweist, weshalb dieses Modell ebenfalls nicht brauchbar ist. Bei dem *Erdos Renyi Modell* besteht die gleiche Problematik wobei hier noch das Problem hinzu kommt, dass ein isolierter Knoten existiert. Ein isolierter Knoten ist im Zusammenhang mit Socialen Netwerken, ein Knoten der keinen Nachbarn besitzt, also Grad 0 aufweist. Dies würde beispielsweise auf Sozial Media, bezogen bedeuten, dass Nutzer*innen auf dieser Plattform angemeldet sind, die keinerlei Verbindungen besitzen. Dies kann durchaus der Fall sein, es ist aber sehr unwahrscheinlich, dass Menschen auf solchen Plattformen angemeldet sind und keinerlei Freunde haben oder andere Nutzer*innen. Schließlich kommt bei den oberen zwei Modellen noch dazu, dass bereits visuell betrachtet kaum bis keine Cliquen und auch keine Brücke auffindbar sind. Deshalb muss auch bei diesem Modell kritisch hinterfragt werden, ob es sich bei den Graphen um ein *typisches soziales Netzwerke* handelt. Deshalb liegt nahe, dass

Anpassungen durchführen werden müssen.

Eine mögliche Optimierung wird erzielt, indem von den Random Graphen-Methoden, die im vorherigen Abschnitt eingeführt wurden, abweichen. Eine weitere Überlegung, um eine Optimierung zu erzielen ist, alle Formeln selbständig zu implementieren und nicht die bereits vordefinierten Funktionen zu verwenden. Zum einen sind diese vordefinierten Funktionen intransparent und daher auch fehleranfälliger, aber auch der Zugriff auf diese ist nicht ganz einfach. Zudem wird durch die eigenständige Implementierung der Zentralitäten ein noch besseres Verständnis der Formeln dieser gewährleistet. Für die Generierung eines *sozialen Netzwerks* wird unter anderem eine Methode benötigt, die einzelne zufällige Graphen erstellt.

Algorithm 1 Random Adjazenzmatrix

```
1:  procedure RANDOM ADJACENCY MATRIX
2:      matrix ← zufällige Matrix der Größe (n,n) zufällig befüllt mit Werten zwischen 0 und 1
3:      for i liegt in der Matrix do
4:          befülle die Diagonale der Matrix mit 1.
5:      for i und k liegen in der Matrix do
6:          setze die Wahrscheinlichkeit auf einen zufälligen Wert zwischen 0 und 1.
7:          if Matrix an der Stelle [i][k] größer als die Wahrscheinlichkeit ist then
8:              setze die Matrix an dieser Stelle auf 0
9:          else
10:             setze diese Stelle auf 1
11:     for i liegt in der Matrix do
12:         was für Matrix an der Stelle [i][j] gilt, muss auch für [j][i] gelten.
13:         RETURN Matrix
```

Dieser Algorithmus erstellt uns zufällige Matrizen, die aber erst noch zu einer großen Matrix zusammengefügt werden müssen. Hierfür benötigt man eine Methode, wie den *Graph appender*. Der Algorithmus dieser Methode soll wie folgt aussehen.

Algorithm 2 alle Subgraphen zu einer Liste zusammenführen

```
1:  procedure GRAPH APPENDER
2:      graphs ← leeres Array
3:      for i zwischen 1 und der Anzahl an Subgraphen / Matrizen do
4:          k ← zufälliger integer, der die Größe des Subgraphen definiert
5:          p ← zufälliger double zwischen 0 und 1 für die Wahrscheinlichkeit
6:          goTo Algorithm 1 mit den übergebenen Werten k und p
7:          füge random Matrix in graphs ein und  RETURN graphs
```

Die einzelne Matrizen werden der Liste hinten angefügt. Nachdem nun eine Liste mit vielen zufällig erzeugten Matrizen generiert ist, fehlt lediglich eine Methode, um die Graphen zusammenzuführen und sicherzustellen, dass die Teilgraphen miteinander verbunden sind. Der Algorithmus sieht hierfür wie folgt aus:

Algorithm 3 Graphs zusammenführen

```
1: procedure UNITE GRAPHS
2:     if länge der Liste graphs aus nur einem Element besteht then
3:         gebe graphs zurück.
4:     dimesion ← 0
5:     big graph ← Graph mit Nullen befüllt
6:     for i zwischen o und der Länge von graphs do
7:         Variable a ← zufälliger integer zwischen 0 und Länge von graphs
8:         Variable b ← zufälliger integer zwischen 0 und Länge von graphs
9:         for j und k zwischen o und graphs do
10:            l ← summierte Länge von Graphs bis zur Stelle i
11:            big graph an der Stelle [(l+j)][(l+k)] ← graph[j][k]
12:            big graph an der Stelle [(l+k)][(l+j)] ← graph[k][j]
13:            big graph an der Stelle [(l+a)][(l+b+graphs Länge an [i]) modulo der Dim] ← 1
14:            big graph an der Stelle [(l+b+graphs Länge an [i]) modulo Dim)][(l+a)] ← 1
15: nun wird der Knoten mit der höchsten Gradzentralität gesucht, um die einzelnen Subgraphen
    miteinander zu verbinden. Dies machen geschieht wie folgt
16:
17:    counter 1 ← 0
18:    counter 2 ← 0
19:    Knoten ← 0
20:    for i und j zwischen o und der Länge von graphs do
21:        if graphs an der Stelle [i][j] ungleich o then
22:            counter 1 ← erhöhe um 1
23:            if counter 1 größer counter 2 then
24:                counter 1 ← counter 2
25:                Knoten ← i
       RETURN Knoten
```

Jetzt ist ein großer Graph generiert, bestehend aus vielen zufälligen kleinen Graphen, welche durch den Knoten mit den meisten ein- und ausgehenden Kanten mit einem weiteren Subgraphen verbunden sind. Nach weiteren Überlegungen ist zusätzlich die Idee entstanden eine Methode zu schreiben, die sicherstellt, dass der generierte Graph aus einer bestimmten Anzahl an Cliquen besteht. Mit diesem zusätzlich Faktor soll sichergestellt werden, dass der generierte Graph mehr Kanten besitzt als davor und die Cluster, in der Visualisierung, schöne Gruppierungen aufweisen. Der Cliquen-Methode soll hierfür eine fixe Zahl n übergeben und zusätzlich sichergestellt werden, dass stetig neue Graphen generiert werden müssen, bis die Anzahl an Cliquen genau der fixen Zahl n entspricht. Durch die Methoden 1, 2 und 3 entsteht schließlich folgender Graph:

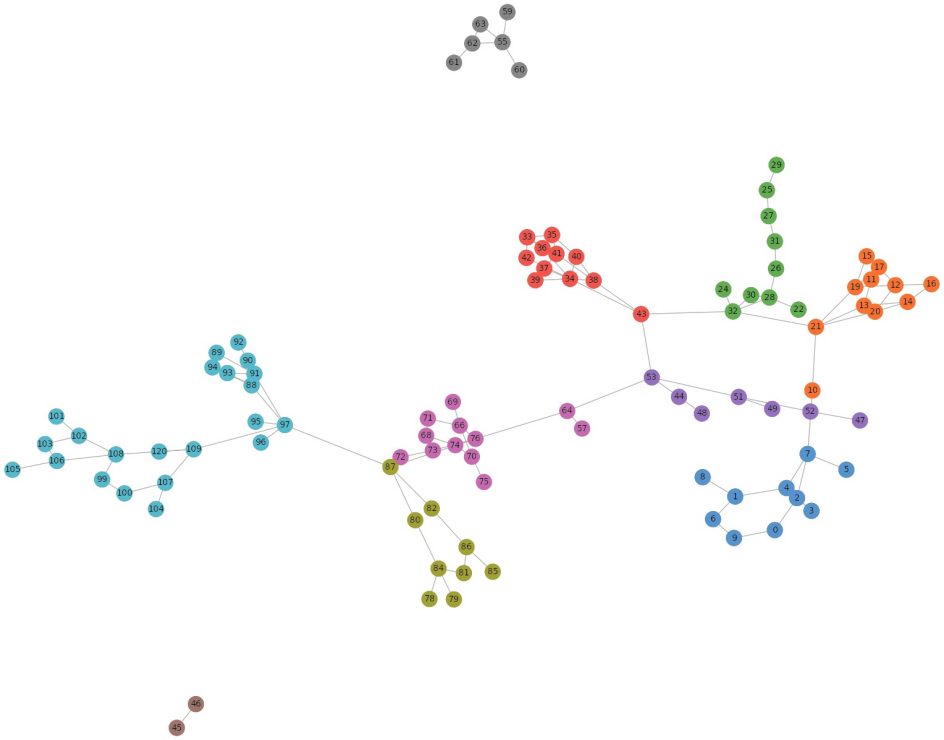


Abbildung 4.3: Random Soziale Graphen mit den höchsten Gradzentralitäts-Knoten als Verbindung

Nachdem der Plot 4.3 durchaus *Sozialen Netzwerken* ähnelt und die Werte der Berechnungen ebenfalls richtig erscheinen, muss noch eine weitere Optimierung durchführen. Bei einer genaueren Betrachtung der Abbildung fällt auf, dass die Teilgraphen wenige Verbindungen, uns bekannte Brücken, untereinander aufweisen. Dies liegt an der Idee von 3, den Knoten mit der höchsten Gradzentralität zu wählen und diesen dann mit einer beliebigen weiteren Gruppe zu verbinden. Doch in tatsächlich ist ein solches Phänomen sehr unwahrscheinlich. Denn dies würde beispielsweise heißen, dass an der Universität Ulm alle Student(en)*innen der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie untereinander in einer Weise miteinander verbunden sind, jedoch nur die Professor(en)*innen, welche die höchste Gradzentralität aufweisen, mit eine*m/r weiteren Professor*in einer anderen Fakultät verbunden sind. Dies ist aber nicht realistisch wenn bedacht wird, dass auch beispielsweise Student(en)*innen der Fakultät für Mathematik und Wirtschaftswissenschaften durchaus Kontakte zu der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie haben können oder auch mit den jeweiligen Professor(en)*innen. Dementsprechend muss diese Eigenschaft ebenfalls in der Implementierung berücksichtigt werden. Das kann gewährleistet werden, indem jedem Knoten eine zufällige Wahrscheinlichkeit zugeschrieben wird, die angibt, ob eine Kante zwischen den Subgraphen existiert. Hierfür wird der Algorithmus 3 ab Zeile *17* ersetzt zu:

Algorithm 4 Verbindung Subgraphen

```
1: procedure CONNECTION SUBGRAPHS
2:     prob ← zufällige Zahl, die sehr klein ist bis 0.00001
3:     for i und j liegen in der Matrix big graph do
4:         befülle die Diagonale der Matrix mit 1.
5:     for i und k liegen in der Matrix do
6:         variable ← zufällige Zahl zwischen 0 und 1
7:         if variable kleiner prob then
8:             setze big graph [i][j] auf 1
9:     RETURN big graph
```

Mit 4 kann sichergestellt werden, dass die Subgraphen vermehrt miteinander verbunden sind und nicht von dem Knoten mit der höchsten Gradzentralität abhängen.

4.2 DIE ANALYSE DES GENERIERTEN GRAPHEN

Mit den Überlegungen aus 4.2 und den dort erklärten Methoden, lassen sich schließlich möglicherweise *typische soziale Netzwerke* realisieren. Um zu beweisen, dass es sich tatsächlich um ein solches Netzwerk handelt, soll ein neues generieren und eine Analyse darauf durchführen. Ziel ist es, zu zeigen, dass die mit dem Generator erzeugten Graphen tatsächlich näherungsweise *sozialen Netzwerken* entsprechen.

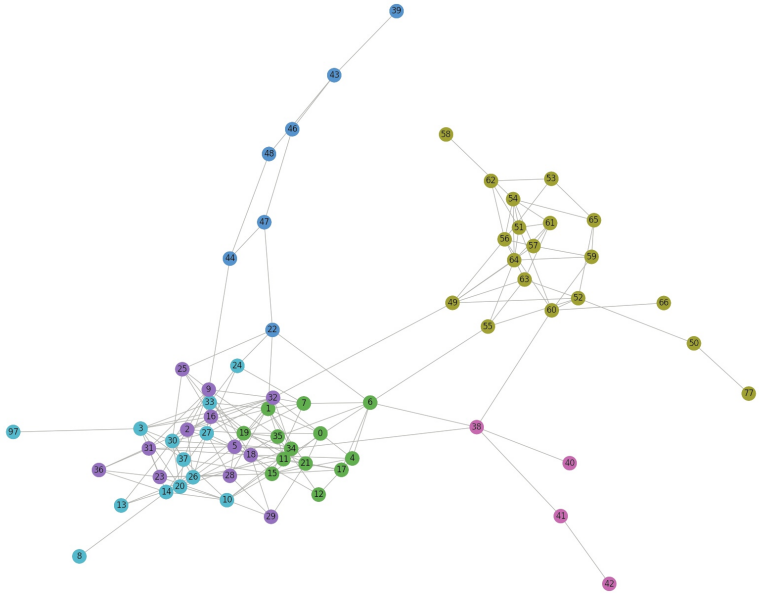


Abbildung 4.4: Random soziales Netzwerk mit realistischeren Verbindungen

Bei der visuellen Betrachtung des Plots 4.4 ähnelt die Struktur auf jeden Fall der, eines sozialen Netzwerks. Doch um eine fundierte Aussagen treffen zu können, müssen auch die Zentralitäten genauer analysiert werden. Hierfür verwendet man folgende Tabelle:

Tabelle 4.1: Werte oberer Graph

Knoten	Grad-Zentr.	Nähe-Zentr.	Between-Zentr.	Knoten	Grad-Zentr.	Nähe-Zentr.	Between-Zentr.
1	0.149254	0.389535	0.0429244	38	0.0746269	0.36612	0.154688
2	0.134328	0.370166	0.0366434	41	0.0298507	0.271255	0.0298507
3	0.119403	0.350785	0.0516569	43	0.0447761	0.198813	0.030303
5	0.119403	0.378531	0.0341306	44	0.0447761	0.295154	0.0773717
6	0.119403	0.385057	0.145038	46	0.0298507	0.219672	0.0205638
7	0.0895522	0.358289	0.0208983	47	0.0447761	0.27459	0.0520902
10	0.119403	0.341837	0.0240985	48	0.0298507	0.232639	0.0373285
11	0.104478	0.360215	0.0212421	49	0.0895522	0.36413	0.221288
14	0.119403	0.3350	0.0454434	50	0.0298507	0.241877	0.0298507
18	0.134328	0.340102	0.0283754	52	0.0895522	0.314554	0.0885577
22	0.0746269	0.348958	0.0740623	54	0.104478	0.254753	0.0327816
27	0.119403	0.360215	0.0342121	55	0.0597015	0.325243	0.0670173
30	0.149254	0.348958	0.0412278	56	0.104478	0.303167	0.0672381
32	0.179104	0.435065	0.266448	57	0.0746269	0.290043	0.0213757
34	0.134328	0.394118	0.112543	60	0.0895522	0.313084	0.0903114
35	0.104478	0.362162	0.0290967	64	0.0895522	0.304545	0.0530434

Bei dieser Tabelle handelt es sich um die **32** wichtigsten Knoten. Die Anzahl der Knoten in der Tabelle 4.2 ist rein zufällig gewählt und hat keine Bedeutung. Alle Knoten die eine geringere **Betweenness-Centrality** kleineren als **0.02** aufweisen, sind außen vor gelassen. Auch wurde dabei die Grenze rein zufällig gewählt. Bei diesem Grenzwert handelt es sich um einen guten Mittelwert. Es sollen weder zu wenig, noch zu viele Knoten betrachten. Bei der Grad-Zentralität aus der Tabelle 4.2 sehen wir, dass die meisten Knoten einen Wert höher als **0.1** aufweisen. Zudem weisen einige, wenige Knoten eine Grad-Zentralität höher als **0.13** auf. Genau genommen handelt es sich hier um die Knoten **1** mit einem Wert von **0.149254**, den Knoten **2** mit dem Wert **0.134328**, Knoten **18** mit dem Wert **0.134328**, dann Knoten **30** mit einer Zentralität von **0.149254**, zudem um den Knoten **32** mit dem höchsten Wert von **0.179104** und schließlich Knoten **34** mit einer Grad-Zentralität von **0.134328**. All diese aufgezählten Knoten sind zentral wichtig für den Graphen und befinden sich höchstwahrscheinlich im Zentrum des Graphen 4.4. Wird nun die Abbildung visuell betrachtet, kann diese Behauptung teilweise bestätigt werden, denn die Knoten fallen direkt auf. Hohe Zentralitätswerte bei einem Knoten sagen aus, dass es sich beispielsweise im realen Leben um eine vermutlich sehr berühmte / bekannte Person handeln wird. Die Annahme besteht, dass es ein Star, ein Influenzer oder eine, auf weitere Arten bekannte Person ist. Doch ebenso liegt nahe, dass die Person lediglich viele andere Personen kennt, oder von vielen anderen Personen gekannt wird. Doch nicht nur die Grad-Zentralität spielt für uns und die Analyse in dieser Arbeit eine zentrale Rolle. Im Weiteren betrachten wird die **Nähe-Zentralität** analysiert, doch um auch bei diesem Aspekt nicht alle **32** Werte aufzuzählen, werden im Folgenden nur Knoten betrachtet, die einen Wert höher als **0.37** aufweisen. Hierzu zählen der Knoten **1** mit einem Wert von **0.389535**, Konten **2** mit dem Wert **0.370166**, zudem Knoten **5** mit dem Wert **0.378531**, zusätzlich Knoten **6** mit der Zentralität

0.385057, und schließlich die Knoten **32** mit dem höchsten Wert **0.435065** und **34** mit der Zentralität von **0.394118**. Je höher die Werte sind, so ist aus dem ersten Teil der Arbeit bekannt, desto näher Befinden sich diese Knoten zu weiteren bzw. weisen die durchschnittlich kürzesten Wege auf. Wird der Graph 4.4 anhand dieser Information betrachtet und werden so die Knoten mit der höchsten **Nähe-Zentralität** gesucht, ist direkt visuell klar, dass sich diese im gleichen Bereich befinden, wie die Knoten mit der höchsten **Grad-Zentralität**. Doch bestätigt der Plot die Vermutung nicht eindeutig, da es teilweise nicht ideal zu erkennen ist, ob die Kanten zum Knoten verlaufen oder an diesem vorbei, weshalb die ausschließlich visuelle Betrachtung eines Graphen oftmals nicht ausreichend ist. Zudem besteht auch die Möglichkeit, dass die Formeln nicht korrekt implementiert wurden. Wobei händisches Nachrechnen, bei durchaus kleineren Plots, die Korrektheit dieser bewiesen hat, weshalb diese Vermutung in Klammern gesetzt wird und es daher tatsächlich an der nicht eindeutigen visuellen Darstellung liegen kann. Die letzte zu untersuchenden Zentralität ist die **Betweenness-Zentralität**. Auch hier betrachtet man wieder die Knoten mit den höchsten Werten, und um nicht alle **32** Werte aufzuzählen, werden erneut nur Knoten mit einem Wert höher als **0.09** ausgewählt. Diese Voraussetzung erfüllen neben dem Knoten **6** mit dem Wert **0.145038** die Knoten **32** mit der höchsten Zentralität von **0.266448** und **34** mit einem Wert von **0.112543**, außerdem der Knoten **38** mit der Zentralität von **0.154688**, zudem der Knoten **49** mit dem Wert **0.221288** und schließlich der Knoten **60** mit dem Wert **0.0903114**. Das bedeutet für unseren Graphen 4.4, dass die kürzesten Wege anteilsmäßig am öftesten über diese genannten Knoten verlaufen. Zudem kann vermuten werden, dass es sich bei diesen Knoten um *Brücken* handelt, was der Behauptung aus dem vorherigen Kapitel, dass es sich bei hohen **Zwischen-Zentralitäten / Betweenness-Zentralitäten** um *Brücken* handelt, bestätigen würde. Bei der erneuten Betrachtung des Graphen, erkennt man eine zum Teil bekannte Eigenschaft, dass sich die Punkte im grün, lila, hellblau verschmolzenen, links unten zentrierten, drei Teilgraphen befinden. Doch kommt bei der **Betweenness-Zentralität** hinzu, dass sich die Knoten **49 und 60** auch im gelbgrünen, rechts oben liegenden, Teilgraphen befinden. Außerdem ist auf den ersten Blick zu erkennen, dass es sich bei den sechs Knoten in 4.4 visuell betrachtet, tatsächlich um Brücken handelt. Im Allgemeinen ist es eindeutig, dass die Werte gut zu den bisher generierten Plots passen und es sehr wahrscheinlich ist, dass die Annahme korrekt ist und die Knoten tatsächlich am häufigsten bei allen kürzesten Wegen durchlaufen werden und tatsächlich Brücken sind. Leider ist die Anzahl an *Cliquen* aus der Abbildung 4.4 nicht mehr exakt bekannt, da erst nach der weiteren Optimierung des Codes Rücksicht drauf genommen wurde. Daher wird an dieser Stelle nur die Vermutung aufgestellt, dass alle Knoten aus der Tabelle 4.2 Teil von Cliquen sind. Wie viele es jedoch genau sind lässt sich an dieser Stelle leider nur wage vermuten. Im nächsten Abschnitt wird jedoch die Betrachtung der Cliquen des Graphen mit einbezogen. Weitere Zentralitätswerte und Eigenschaften des Graphen werden an dieser Stelle nicht betrachtet. Nachdem alle Kriterien überprüft sind und erfolgreich festgestellt wurde, dass dieser Graph einem **sozialen Netzwerk** ähnelt, wird noch untersucht, ob nicht womöglich zufällig ein passender Graph erhalten wurde. Deshalb wird noch ein weiteres Kriterium für die Zentralitäten überlegt und anschließend untersucht.

4.3 DIE VERTEILUNG DER ZENTRALITÄTEN

Nachdem im vorherigen Kapitel die Generierung eines **sozialen Netzwerkes** und die Analyse durchgeführt wurde, spielt im Folgenden die Verteilung der Zentralitätswerte eine wichtige Rolle. Im Laufe der Arbeit ist aufgefallen, dass sich die Werte der Zenralitäten oftmals in ähnlichen Zahlenbereichen befinden. An dieser Stelle entsteht die Frage, wie diese Werte verteilt sind und ob die Verteilung einer mathematischen Wahrscheinlichkeitsverteilung entspricht. Das heißt, im Konkreten, es wird der Frage nachgegangen, ob alle Zentralitätswerte sozialer Netzwerke ähnliche Verteilungen nachweisen. Wenn sich die Vermutung diesbezüglich bestätigt, können andere soziale Netzwerke anhand dieses Kriteriums verglichen werden. Da der Graph 4.4 ein zufällig, einmalig erzeugter Graph ist, muss ein neuer Graphen in unserem Generator erzeugt werden, um die Verteilung der Zentralitäten zu betrachten. Dies wird sich nicht auf die Untersuchung auswirken, denn die Verteilungen der Zentralitäten unserer Graphen sollte grob gleich oder zumindest ähnlich sein. Bei der erneuten Generierung entstehen nun folgende Graphen und die zugehörige Verteilung der **Gradzentralität**:

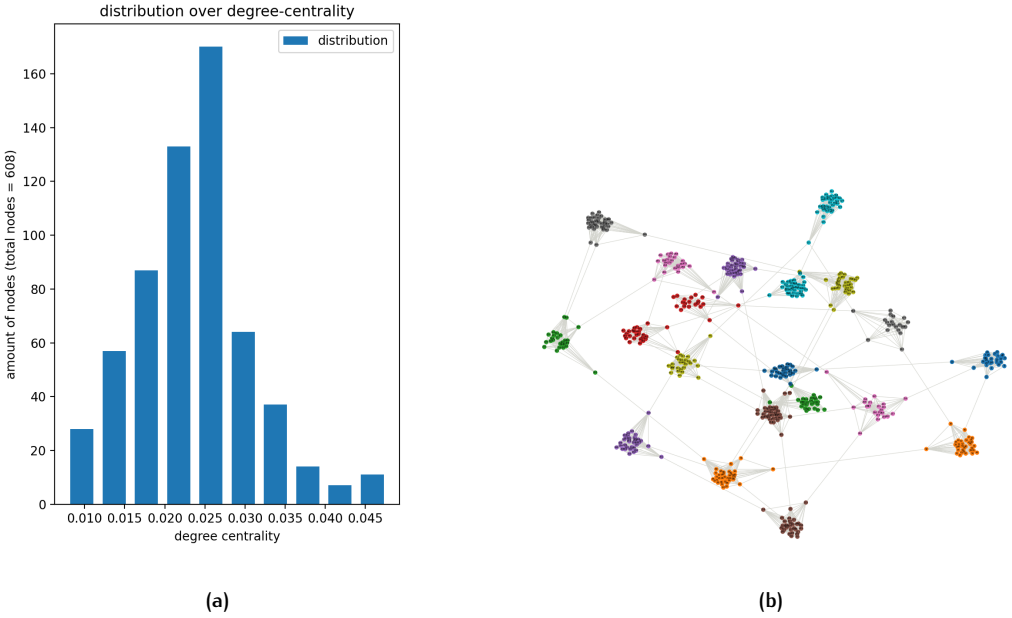


Abbildung 4.5: Verteilung der Grad-Zentralität des Graphen (b)

Es wird ersichtlich, dass die **Gradzentralität** normal- beziehunsgweise gaußverteilt ist. Natürlich ist zu erwähnen, dass keine perfekte Gauß-Verteilung zu sehen ist, sondern eine etwas nach links verschobene Verteilung. Was die möglichen Gründe dafür sind, werden später betrachtet und korrigiert. Nun wird untersuchen, ob sich die Eigenschaft, der normalverteilten Zentralitäten für die **Nähe-** und **Betweenness-Zentralität** ebenfalls bestätigen lässt. Um zusätzlich zu beweisen, dass es sich bei der Gauß-Verteilung der Werte nicht um einen Zufall handelt, wird ein neuer sozialer Graph generiert und die Verteilung der **Grad-**, **Nähe-**, **Betweenness-** und **Eigenvektor-Zentralität** untersucht. Hierbei sind vor allem die Frage, ob die Verteilung

einer tatsächlichen Normalverteilung entspricht und falls ja, warum dies der Fall ist, essentiell. Ansonsten wird die Frage gestellt, warum es keiner Normalverteilung entspricht und ob es möglich ist, den Graphen zu verändern um eine solche zu erzielen. Bei der erneuten Generierung entstehen schließlich folgende Plots:

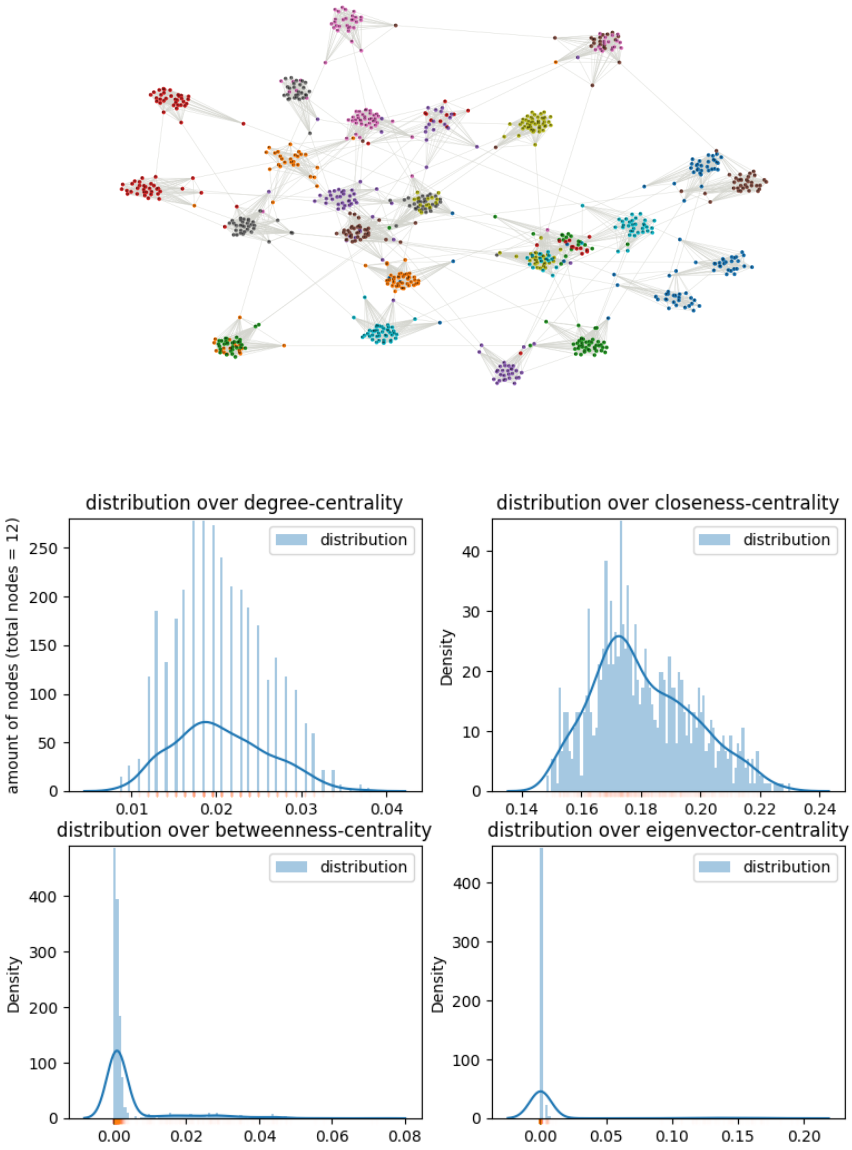


Abbildung 4.6: Random soziales Netzwerk mit realistischeren Verbindungen

In der Abbildung 4.6 sieht man nun die Verteilungen der Zentralitäten von dem, sich darunter befindenden, sozialen Netzwerks. Die Tabelle mit den Zentralitäts-Werten des Netzwerks

befindet sich als Datei in [20]. Oben links befindet sich die Verteilung der **Grad-Zentralität**, welche wie bereits oben festgestellt, nicht exakt normalverteilt ist, aber Ähnlichkeiten zu erkennen sind. Vor allem auffällig ist, dass der Balken bei **0.012** vergleichsweise sehr hoch ist. Über **50** Knoten weisen diesen Wert auf. Danach geht der darauf folgende Balken nochmals zurück, denn nur noch etwas über **25** Knoten haben eine Zentralität von circa **0.03**. Jedoch war zu erwarten, dass sich das Balkendiagramm typischerweise symmetrisch verhält, doch das Gegenteil tritt ein. Über **150** Knoten weisen eine Zentralität von **0.0125** auf, daher sollten auch ebenso viele den Wert **0.025** besitzen. Hingegen ist positiv hervorzuheben, dass genau **ein Peak** erreicht wurde, wie auch zu erwarten war. Außerdem sind alle Balken vor dem Peak kontinuierlich aufsteigend und nach dem Peak kontinuierlich absteigend. Doch lediglich eine Unstimmigkeit sticht hier heraus, bei dem Zentralitätswert von **0.0357** den etwas unter **25** Knoten besitzen. Fraglich ist hier, warum der Balken erneut höher ist als sein Vorgänger. Denn im Regelfall sollten maximal ein bis drei Knoten gefunden werden, die diesen Wert aufweisen. Doch im Allgemeinen weist der Plot genau die Eigenschaft nach, die auch zu erwarten ist, nämlich dass die **Grad-Zentralität** annähernd normalverteilt ist.

Das Balkendiagramm der **Nähe-Zentralität bzw. closeness-centrality** weist einen ähnlichen Verlauf auf. Ein bereits erwartetes Peak und weitere Balken, die im linken Bereich sehr schnell zum Peak hin ansteigen und rechts vom Peak vergleichsweise langsam abflachen. Sehr analog zu dem Balkendiagramm der **Grad-Zentralität**. Auffällig ist erneut, dass der letzte Balken wider Erwartens höher ist als der Balken davor. Die Vermutung liegt nahe, dass es sich hier um einen Zufall handelt. Im Rahmen dieser Arbeit wurden viele Generierungen durchgeführt und in [20] befinden sich ebenso weitere Balkendiagramme und Soziale Netzwerke. Die Aussage, welche auf jeden Fall getroffen werden kann ist, dass falls es zu Unstimmigkeiten kommt, welche stets an anderen Stellen auftreten und nicht immer denselben Balken betreffen. Diese Aussage können jedoch nur für die Verteilung der **Grad-** und **Nähe-Zentralitäten** getroffen werden. Jedoch kann ebenso angenommen werden, dass je größer der Graph ist, umso eher sind diese Zentralitäten normalverteilt. Was daran liegt, dass dieser dann mehr Knoten besitzt und diese irgendwann zwangsläufig eine Regelmäßigkeit aufzeigen, da Zahlen in der Mathematik prinzipiell nicht *zufällig* sind. Grob angedeutet, kann die Existenz einer Kante als Binomialverteilung interpretiert werden und diese konvergiert mathematisch gesehen bei einer sehr großen Stichprobe (Anzahl an Knoten in unserem Fall) gegen eine Normal- bzw Gaußverteilung. Doch betrachtet man auch noch die zwei unteren Balkendiagramme an, wird die Behauptung verworfen.

Bei der **Betweenness-** und **Eigenvektor-Zentralität** wird die angenommene Regelmäßigkeit nicht bestätigt. Zum einen weisen die Balken wenige unterschiedliche Werte auf, die teilweise kaum zu unterscheiden sind. Was zudem auffällt, sind die Ausschläge der vorderen Balken. Was zunächst verwunderlich erscheint, ist mit einer simplen Erklärung begründet. Die **Closeness-Zentralität** gibt bekanntlich an, wie oft ein Knoten anteilsmäßig bei der Suche nach dem kürzesten Weg durch einen Graphen benutzt wird. Der Ausschlag ist daher die Folge davon, wenn viele kürzeste Wege stets über die gleichen Knoten verlaufen. Das heißt, es existieren keine bis wenige Alternativen und so verlaufen die kürzesten Wege von beispielsweise **Knoten 1** zu einem weiteren Knoten stets über gleiche beziehungsweise ähnliche Knoten. Bei der **Eigenvektor-Zentralität** wurden zwar die gleiche Beobachtung gemacht, doch sagt diese hier etwas anderes aus. Diese Zentralität gibt eine Einschätzung der Wichtigkeit des Knotens, im Bezug auf seine Nachbarn an, was bezogen auf die Balkendiagramm heißt, dass viele Knoten

in diesen Graphen wichtig sind mit Einbeziehung der Nachbarn. Wobei auch vermuten werden darf, dass dies mit der hohen Anzahl an Konten mit höher **Betweenness-Zentralität** zusammenhängt. Das heißt im Umkehrschluss wiederum, dass mehr Cliquen im Graph enthalten sind. Tatsächlich sind es **935 Knoten**, **8952 Kanten** und **10301** Cliquen mit der maximalen Größe von acht Knoten in der Clique. Die komplette Analyse des sozialen Netzwerks in 4.6 befindet sich in [20], da bereits eine ausführliche Analyse in dieser Arbeit durchgeführt wurde. Danach zu Urteilen handelt es sich bei dem Netzwerk um ein typisches **soziales Netzwerk**.

4.4 KURZES RECAP

Nachdem zunächst überlegt wurde, wie soziale Netzwerke generiert werden, sind auch gleichzeitig die Probleme der Generierung aufgefallen. Daher wurde der Code fortlaufend optimiert, ein soziales Netzwerk erstellt und danach eine soziale Netzwerk Analyse durchgeführt. Dadurch erhält man die Bestätigung, dass es die Anforderungen an ein soziales Netzwerk erfüllt. Danach ist zudem aufgefallen, dass die Zentralitäten regelmäßig sind und eine Normalverteilung nachgewiesen werden kann. Doch muss im Folgenden die Frage beantwortet werden, wie die Verteilung der Zentralitäten bei anderen, bereits analysierten Netzwerken aussieht.

5 DER SOZIALE NETZWERKE VERGLEICH

Im vorherigen Teil der Arbeit haben wir uns damit beschäftigt, wie soziale Netzwerke so gut und realitätsnah wie möglich konstruiert werden können. Wir haben Analysen durchgeführt und festgestellt, dass die Werte unserer **Grad-** und **Nähe-Zentralität** näherungsweise normalverteilt sind. Daher liegt es nahe, weitere sozialen Netzwerke und ihre Analysen zum Vergleich heranzuziehen. Leitfragen sind hierbei, was zu erwarten ist, ob die Ergebnisse den Erwartungen entsprechen oder sogar widersprechen und warum dies der Fall ist. Zusätzlich möchten wir optimalerweise eine Möglichkeit erarbeitet, wie wir unsere Graphen bzw. die Generierung angepasst könnten um möglicherweise noch bessere Graphen zu erhalten, die die sozialen Netzwerken noch mehr ähneln.

5.1 DER DATENSATZ UND DIE ANALYSE

Auf der Suche nach vergleichbaren sozialen Netzwerken, beziehungsweise Datensätzen, ist die Suche scheinbar endlos. Auf vielen Webseiten sind große Datensätze für alle Nutzer*innen zugänglich. Meistens als **CSV** Datei, welche ideal zur Erstellung von Graphen mit unserem Generator geeignet sind. In diesem Teil der Arbeit betrachten wir mehrere Datensätze. Natürlich aufgrund der Tatsache, dass sie spannend sind aber auch um mehrere Vergleichswerte zu haben. Starten wir zunächst mit den Daten [11] von unserem **Game of Thrones** Plot 3.3. Da bereits die Analyse der **Zentralitäten** und die generelle visuelle Analyse des Graphen durchgeführt ist, reicht nun lediglich die Verteilung der Zentralitäten zu betrachten. Die Tabelle mit den Werten der Zentralitätsberechnungen befinden sich erneut in [20]. Nachdem der Datensatz als **CSV** Datei in dem Generator eingelesen und anschließend geplottet wurde, wird folgender Graph konstruiert:

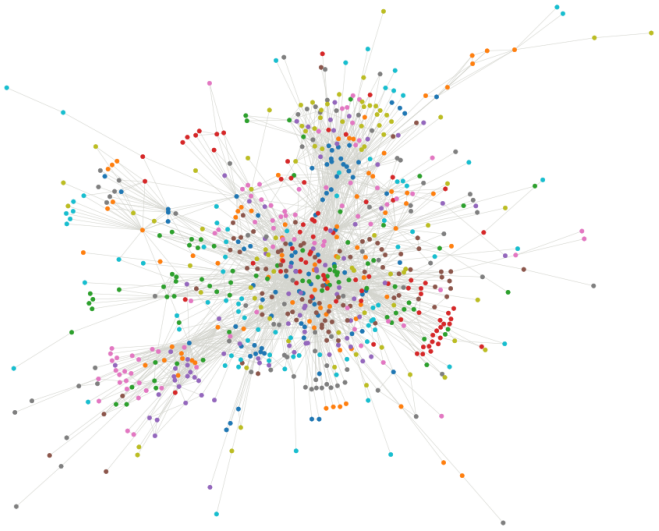


Abbildung 5.1: Game of Thrones Graph 2.0,
selbst erstellt

Dieser Plot bleibt beabsichtigt unkommentiert, da er lediglich zur Argumentation für die Verteilung der Zentralitäten benötigt wird und daher die visuelle Form des Graphen nur von zweitrangiger Bedeutung für diese Arbeit ist. Zudem ist zu vermerken, dass der eigentliche Datensatz gewichtet ist, und die bisher generierten Graph daher bereits schon visuell nicht dem Graphen aus 3.3 ähnelt. Jedoch ist es sinnvoll die Gewichte außen vor zu lassen, da in dieser Arbeit ausschließlich ungewichtete Graphen nachbildet beziehungsweise behandelt werden. Nachdem die Daten des Graphen 5.1 eingelesen, die Zentralitäten berechnet sind und anschließend die Balkengraphen erstellt wurden, ist folgender Plot entstanden:

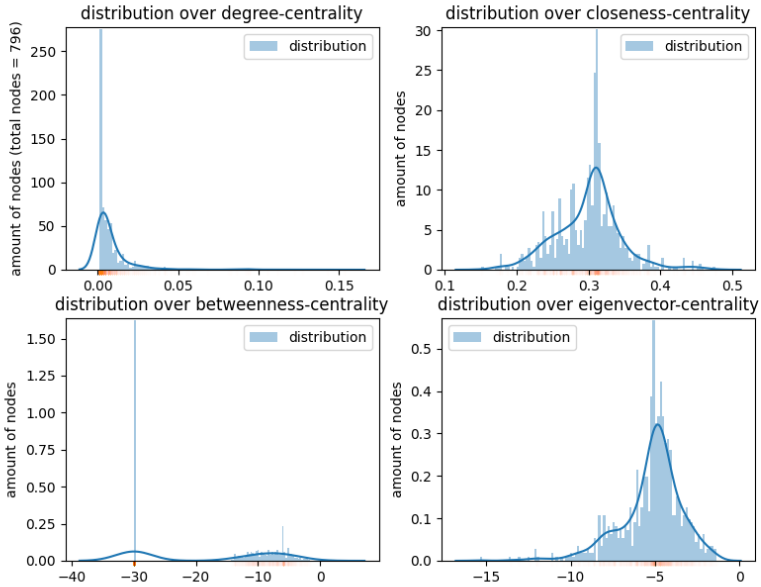


Abbildung 5.2: Game of Thrones Verteilung der Zentralitäten

Auf den ersten Blick wird bereits klar, dass andere Ergebnisse erwartet wurden. Einzig die Verteilung der **Nähe-Zentralität** ähnelt der erwarteten Normalverteilung. Die **Betweenness-** und **Eigenvektor-Zentralität** hingegen ähneln zwar nicht exakt dem, was in 4.6 herausgekommen ist, aber zieht auf jeden Fall Parallelen. Denn beide haben einen Ausschlag von mindestens einem Balken, was bereits im vorherigen Kapitel damit begründet wurde, dass es die Folge von vielen kürzesten Wegen ist, die stets über die gleichen Knoten verlaufen. Daher keine Alternativen im Graph existieren. Die **Grad-Zentralität** hingegen ist tatsächlich verwunderlich. Sie ähnelt keinesfalls der Normalverteilung aber sieht sehr nach einer Exponentialverteilugn aus. Der Ausschlag der Balken ist hingegen schnell erklärt. Es sind viele Konten, in diesem Fall repräsentierte *Game of Thrones* Charaktere, die alle gleich wichtig für den Graphen sind. Diese Knoten sind daher mit vielen anderen Knoten verbunden, werden also von vielen anderen Charakteren gekannt oder kennen viele andere Charaktere. Im Allgemeinen sind die Balkendiagramme der **Zentralitäten** aus 5.2 leider nicht zufriedenstellend. Der Grund, warum die Ergebnisse stark von unseren Erwartungen abweicht ist vermutlich, dass es sich bei dem Graphen um fiktive Charaktere handelt. Dadurch kann es schnell zu Unstimmigkeiten kommen. Zudem war der Datensatz davor gewichtet, was zu anderen Werten bei der Berechnung der Zentralitäten führen kann. Doch wurde der Datensatz ungewichtet betrachtet, um ihn besser mit den generierten Graphen zu vergleichen, welche ungewichtet sind. Dies kann auf jeden Fall ein plausibler Grund für Unstimmigkeiten sein. Zudem haben ist die Anzahl der geplotteten Balken stark erhöht und so fallen Unstimmigkeiten generell deutlich schneller auf. Dennoch soll die Theorie, dass Zentralitäten normalverteilterteilt sind, nicht verworfen werden und wir betrachten noch einen weiteren Datensatz. Der nächste Datensatz, der aus aus "Kreisen"(oder "Freundeslisten") besteht, ist von Facebook veröffentlicht worden. Die Daten wurden jedoch

vor der Veröffentlichung von Facebook anonymisiert, daher ist lediglich bekannt, dass es sich bei dem Datensatz um politische Interessen handelt. So kann mit dem Datensatz feststellt werden, dass zwei Nutzer die gleiche politische Zugehörigkeit haben, aber nicht, was ihre individuelle politische Zugehörigkeit bedeutet [5]. Nachdem die Daten wieder in eine **.CSV** Datei umgewandelt und anschließend geplottet wurden, ist folgenden Graphen entstanden:

Abbildung 5.3: Facebook Graph

Der Graph ähnelt auf den ersten Blick keinem, der bisher generierten Graphen. Zudem fällt aber sofort auf, dass dieser Graph aus deutlich mehr Knoten besteht, zudem weniger Subgraphen besitzt aber dennoch eine grundsätzlich ähnliche Struktur zu unseren anderen Graphen aufweist. Die Berechnungen der Zentralitäten befinden sich ebenfalls auf Github [20], da es sich um zu viele Werte handelt. Nun interessiert uns jedoch, wie diese Zentralitäten verteilt sind und ob dieser Graph die erwarteten Verteilungen nachweist. Nachdem der Graphe 5.3 durch unsere Methode, welche die Plots über die Verteilungen erstellt, gelaufen ist, sind folgende Diagramme entstanden:

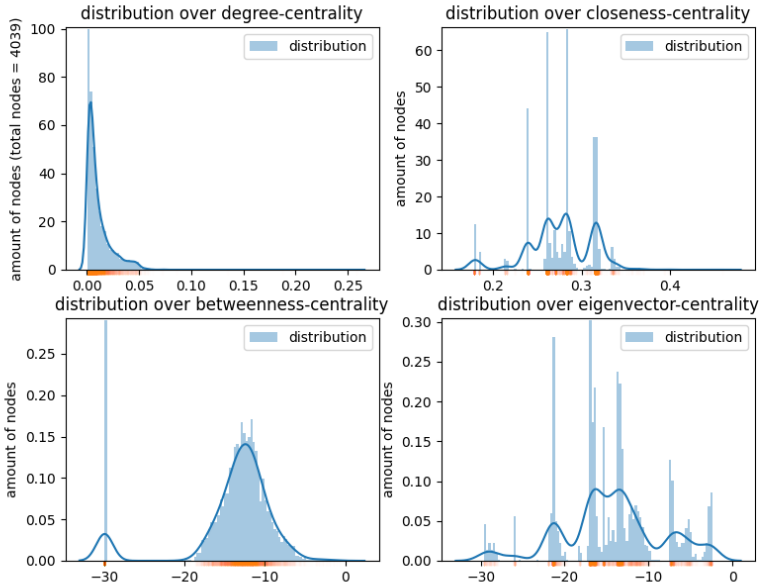


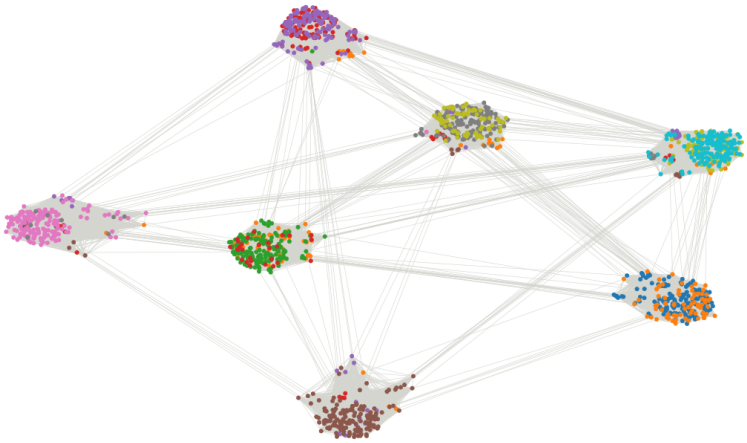
Abbildung 5.4: Facebook Graph Distribution

Sofort fällt auf, dass bei keiner Zentralität eine Normalverteilung erkennbar ist. Die **Grad-Zentralität** fällt aber direkt auf, denn es handelt sich hier um eine Exponentialverteilung. Die anderen Balkendiagramme der **Betweenness-** und **Eigenvektor-Zentralität** ähneln jedoch den Verteilungen aus 4.6. Was zudem auffällig ist, dass die Diagramm stark an die Verteilungen von 5.2 erinnern. Auch wenn diese Ergebnisse sehr ernüchtern scheinen und vor allem das Balkendiagramm der **Nähe-Zentralität** erneut nicht normalverteilt ist, wollen wir uns überlegen, woran dies liegen kann. Bei den anderen Zentarlitäten herrscht eine starke Fluktuation der Balken, wodurch keine Mathematischen Wahrscheinlichkeitsverteilungen erkennbar ist. Visuell fällt jedoch auf, dass der Graph 5.4 verglichen mit dem Plot des Graphen 2 durchaus Parallelen aufweist. Es sind aber deutliche Ansammlungen von Knoten, die auch als Teilgraphen bezeichnet werden können, ersichtlich. Zwischen den Teilgraphen sind, so wie bei 4.4, einige Kanten zu erkennen, die Teilgraphen untereinander verbinden. Natürlich weist der obige Graph deutlich mehr Kanten und Knoten auf als die bisherigen Graphen. Unsere Graphen haben im Schnitt um die **950** Knoten und **8700** Kanten, daher also circa neun mal so viele Kanten wie Knoten. Auch existieren im Schnitt um die **10100** Cliquen, welche maximal acht Knoten groß sind. Bei dem Facebook Graphen 5.3 hingegen **4093** Knoten und **88234** Kanten. Das heißt circa einundzwanzig mal so viele Kanten wie Knoten. Leider ist die Anzahl an Knoten und Kanten des Graphen 5.3 nur durch die Homepage [5] bekannt, denn der Datensatz ist leider zu groß, um die Analyse der Zentralitäten und die Untersuchung auf Cliquen zu Ende zu führen. Daher ist auch die genaue Anzahl an Cliquen dieses Graphen unbekannt, doch kann vermutet werden, dass diese sicherlich deutlich höher sind als bei 4.6, denn es existieren mehr Knoten mit ähnlich hohen Zentralitäten. Schließlich wird noch ein letzter Versuch gestartet, die Kanten und Knoten im Code des Graphen Generators zu erhöhen, um damit die selbe Relation zu erhalten. Dadurch

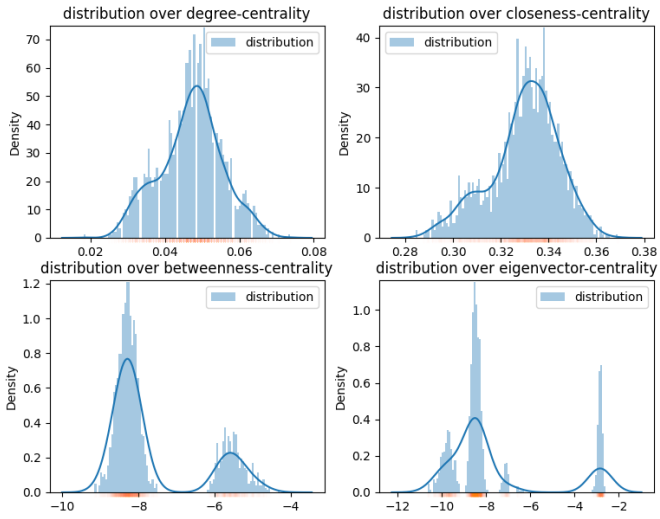
wird womöglich gezeigt, dass alle vier untersuchten Zentralitäten annähern Normalverteilt sind. Was zum einen daran liegt, dass letztendlich Graphen wie 5.1 erhalten sind, die noch viel dichter besetzt sind und leider alle Knoten mit beinahe allen anderen Knoten verbunden sind, was eher untypisch für *soziale Netzwerke* ist. Es darf aber auf jeden Fall festgehalten werden, dass die **Betweenness-** und **Eigenvektor-Zentralität** bei allen untersuchten Datensätzen starke Parallelen zu den Verteilungen des generierten sozialen Netzwerk nachweisen. Jedoch ist nach wie vor die Verteilung der **Gradzentralität** verwunderlich. Daher ist es ratsam, den Code und die damit verbundenen Plot so anzupassen, dass es dem Plot 5.4 ähnelt. Anschließend kann die Verteilungen der Zentralitäten betrachtet und dadurch eine genauere Aussage erzielt werden.

5.2 ANPASSUNG DES GENERIERTEN PLOTS

Nachdem die Verteilungen durchaus verwunderlich ist, wird der nächsten Plot weitestgehend an 5.4 angepasst. An dieser Stelle muss durchaus betont werden, dass es sich bei den bisher generierten Plot keinesfalls um untypische oder falsche soziale Netzwerke handelt. In diesem Abschnitt wird lediglich eine bessere Vergleichsbasis hergestellt. Dies wird ermöglicht, indem zum einen die Anzahl an Cluster auf **sieben** Stück anpasst und die Anzahl der Knoten pro Cluster erhöht wird. Gleichzeitig aber die jeweiligen Größen deutlich mehr variieren lassen und vor allem die Kanten-Menge, also Anzahl an Verbindungen, deutlich erhöhen. Schließlich erhält man folgende Graphen:



(a)



(b)

Abbildung 5.5: Final optimierter Graph

Natürlich ist direkt ersichtlich, ohne die Werte genauer analysiert zu haben, dass keine zu 5.4 identischen Graphen erzeugt werden können. Dies liegt an mehreren Faktoren, die unter anderem im Ausblick erörtern werden. Allgemein sind die, in dieser Arbeit generierten Graphen, auch wenn die Varianz der Graph-Größen best möglichst garantiert wird, auf den ersten Blick ähnlich groß. Jedoch wurde bei der Verteilung der Zwischen- und Eigenvektor Zentralität eine absolute Verbesserung erzielt, indem die Zentralitäten, bevor sie geplottet werden, logarithmiert werden. Dies wurde nachträglich auch bei allen vorherigen Verteilungen

gemacht. Das ermöglicht es, die Verteilung auseinander zu zerren, da sich die Werte stets um **0.0** verteilt haben. Nun fällt auf, dass die Verteilungen dieser zwei Zentralitäten sehr ähnlich sind. Eine mögliche Begründung hierfür kann sein, dass die Eigenwert-Zentralität im Nachhinein betrachtet nicht sonderlich interessant ist. Nehmen wir an, dass x^* ein nicht negativer Eigenvektor ist, mit einem Eintrag für jeden Knoten. Aufsummiert soll dieser *1* ergeben. Dadurch darf er als Wahrscheinlichkeitsverteilung interpretiert werden. Für einen Knoten v kann der Eintrag von x_v^* als Eintrag seiner *Eigenwert-Zentralität* gesehen werden. Tatsächlich kann ein Eigenvektor mit Bezug auf 1 leicht erraten werden.
Sein nun

$$x_v^* = \frac{k_v}{2 \times |E|} \forall v \in V \tag{5.1}$$

Wobei k_v die *Grad-Zentralität* des Knoten **v** ist und |E| die Anzahl an Kanten. Es folgt direkt, dass x_v^* die Bedingungen für x^* erfüllt. Tatsächlich erinnert uns 5.1 an 3.1 das heißt, die *Eigenvektor-Zentralität* entspricht der mit einer Konstanten skalierten *Grad-Zentralität*. Das ist jedoch nur eine Vermutung und wird nicht mehr weiter betrachtet. Schließlich fehlt noch die Einordnung des 5.5 und 5.4 Hier sind leider keine identischen Verteilungen entstanden. Tatsächlich kann dies auch mathematisch begründet werden, denn wenn zwei Zufallsvariablen **X** und **Y** standardnormalverteilt und unabhängig sind, dann wären für Parameter $\lambda = \frac{1}{2}$ die Variablen $X^2 + Y^2$ exponentialverteilt [19]. Doch hat die Optimierung in diesem Kapitel dennoch viel gebracht. Unter anderem konnte bewiesen werden, dass mit wenigen Anpassungen des Codes, annähernd vergleichbare Verteilungen erhalten werden. Um den idealen Vergleich herzustellen, müssen jedoch größere Änderungen am Code vorgenommen werden, was jedoch nicht mehr im Umfang dieser Arbeit liegt.

6 FAZIT UND AUSBLICK

Nachdem in dieser Arbeit ausführlich die Generierung und Analyse sozialer Netzwerk behandelt wurde, werden nun die wichtigsten Erkenntnisse zusammenführen. Die Analyse sozialer Netzwerke besteht aus vielen Faktoren. Es gibt zahlreiche Methoden um eine Analyse durchzuführen und viele charakteristische Merkmale, die bei dieser von Bedeutung sind. In dieser Arbeit wurde gezeigt, dass die Betrachtung der Cliquen und Brücken bei der visuellen Interpretation ausreicht. Bei der Analyse der Daten ist zudem deutlich geworden, dass es besser ist die Verteilungen dieser zu betrachten. Gleichzeitig ist zu beachten, dass soziale Netzwerke unterschiedlichste Thematiken darstellen. Alleine eine kurze Recherche im Internet präsentiert unzählige unterschiedliche Netzwerke. Doch haben diese eine Gemeinsamkeit, sie sind alle auf ihre weise typische soziale Netzwerke. Anhand dieser Arbeit wurde ebenfalls ersichtlich, dass bereits kleine Optimierungen im Code, die Generierung visuell ähnlicher Graphen ermöglicht. Sobald die Graphen gleiche visuelle Grundstrukturen aufweisen, folgen auch starke Gleichheiten in der Verteilung der Zentralitäten. Diese Arbeit lässt einige Punkte offen, die durchaus noch weiter optimiert werden können. Beispielsweise die generierten Plots noch besser an existierende Graphen anpassen, um die Verteilung bestmöglich nachzustellen. Unser Generator bildet lediglich kleine Subgraphen und verbindet diese, wobei aber es auch Knoten geben kann, die sich zwischen den Clustern befinden siehe 5.3. Dies könnte durch eine weitere Methode im Code erzeugt werden. Zudem wäre ein weiterer interessanter Faktor die Dichte in diesen Clustern, ob die Knoten sehr nah beieinander liegen oder weit voneinander entfernt sind. Auch größere Datensätze zu untersuchen und zu vergleichen wäre eine interessante Fortsetzung dieser Arbeit. Oder ebenfalls interessant, ob die Berechnungen der Zentralitäten bereits optimierte Algorithmen sind und ob nicht möglicherweise doch Optimierungspotenzial besteht. All diese Ideen zeigen erneut, wie vielfältig soziale Netzwerke und die Analyse sind und warum sie zahlreiche Wisschenschaftler*innen beschäftigt. Schließlich kann diese Arbeit damit beendet werden, dass es unglaublich vielzählige Methoden zu Analyse von Netzwerken gibt. Welche die geeignetste ist, lässt sich nicht in einem Satz formulieren. Es kommt auf Anzahlen von Kanten und Knoten an, aber auch auf die zu untersuchende Thematik.

LITERATUR

[1] NetworkX Developers. *Graph generators*. 2014-2022. URL: https://networkx.org/documentation/stable/reference/generators.html (besucht am 28. 03. 2022).

[2] Jennifer Golbeck. "Chapter 3 - Network Structure and Measures". In: *Analyzing the Social Web*. Hrsg. von Jennifer Golbeck. Boston: Morgan Kaufmann, 2013, S. 25–44. ISBN: 978-0-12-405531-5. DOI: https://doi.org/10.1016/B978-0-12-405531-5.00003-1. URL: https://www.sciencedirect.com/science/article/pii/B9780124055315000031.

[3] Riddle M. Hanneman R. *Introduction to Social Network Methods (Hanneman)*. University of California, Riverside, 2019. URL: https://math.libretexts.org/@go/page/7645.

[4] Charles Kadushin. "Introduction to Social Network Theory". In: (Jan. 2004).

[5] By Jure Leskovec. *Social circles: Facebook*. 2012. URL: https://snap.stanford.edu/data/ego-Facebook.html (besucht am 28. 03. 2022).

[6] Elbert E N Macau. *A mathematical modeling approach from nonlinear dynamics to complex systems*. Springer, 20198. URL: https://www.worldcat.org/title/mathematical-modeling-approach-from-nonlinear-dynamics-to-complex-systems/oclc/1117866920.

[7] Peter Marsden. "Egocentric and Sociocentric Measures of Network Centrality". In: *Social Networks - SOC NETWORKS* 24 (Okt. 2002), S. 407–422. DOI: 10.1016/S0378-8733(02)00016-3.

[8] Ruchi Nayyar. *Representing Graphs in Data Structures*. Oktober 2017. URL: https://www.mygreatlearning.com/blog/representing-graphs-in-data-structures/ (besucht am 28. 03. 2022).

[9] Christina Newberry. *How to Find and Target Your Social Media Audience (Free Template)*. 2020. URL: https://blog.hootsuite.com/target-market/ (besucht am 28. 03. 2020).

[10] Ioannis Panges. *Social Network Analysis.An Introduction*. GRIN Verlag, 2016. URL: https://www.grin.com/document/371489.

[11] George Pipis. *Social Network Analysis Of Game Of Thrones In NetworkX*. September 2019. URL: https://predictivehacks.com/social-network-analysis-of-game-of-thrones/ (besucht am 28. 03. 2022).

[12] Francisco Rodrigues. "Network Centrality: An Introduction". In: März 2018. ISBN: 978-3-319-78511-0. DOI: 10.1007/978-3-319-78512-7_10.

[13] Britta Ruhnau. "Eigenvector-centrality — a node-centrality?" In: *Social Networks* 22 (Okt. 2000), S. 357–365. DOI: 10.1016/S0378-8733(00)00031-9.

[14] John P. Scott und Peter J. Carrington. *The SAGE Handbook of Social Network Analysis*. Sage Publications Ltd., 2011. ISBN: 1847873952.

[15] Laura Sheble, Kathy Brennan und Barbara Wildemuth. "Social network analysis". In: Jan. 2016, S. 250–339. ISBN: 978-1440839047.

[16] Unknown. *Social Networks*. 2021, February 20. URL: https://socialsci.libretexts.org/@go/page/8043 (besucht am 28. 03. 2022).

[17] Unknown. *Web 2.0 and Social Media*. 2022, March 02. URL: https://mitchell.libguides.com/c.php?g=529360&p=3620303 (besucht am 28. 03. 2022).

[18] Stanley Wasserman und Katherine Faust. *Social network analysis: Methods and applications*. Bd. 8. Cambridge university press, 1994. URL: http://scholar.google.com/scholar.bib?q=info:gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0.

[19] Wikipedia. *Exponentialverteilung*. 2008-2022. URL: https://de.wikipedia.org/wiki/Exponentialverteilung#Beziehung_zur_Normalverteilung (besucht am 20. 04. 2022).

[20] Tanja Zast. *Social Network Analysis*. 2022. URL: https://github.com/TanjaZast/bachelorthesis-sna (besucht am 28. 03. 2022).

ERKLÄRUNG

Hiermit erkläre ich, dass ich die vorliegende Ausarbeitung selbst und ohne Verwendung anderer als der zitierten Quellen und Hilfsmittel verfasst habe. Wörtlich zitierte Sätze oder Satzteile sind als solche kenntlich gemacht; andere Hinweise zur Aussage und zum Umfang sind durch vollständige Angaben zu den betreffenden Publikationen gekennzeichnet. Die Ausarbeitung wurde in gleicher oder ähnlicher Form keiner Prüfungsstelle vorgelegt und ist nicht veröffentlicht worden. Diese Arbeit wurde noch nicht, auch nicht teilweise, in einer anderen Prüfung oder als Lehrveranstaltungsleistung verwendet.

Ulm, April 2022

Tanja Zast