

TANJA & ZAST

METHODEN DER ANALYSE VON SOZIALEN NETZWERKEN



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

Kurze Zusammenfassung des Inhaltes in deutscher Sprache...

INHALTSVERZEICHNIS

I	EINFÜHRUNG IN DIE THEORIE	
1	EINLEITUNG	2
1.1	Zielsetzung	2
2	EINFÜHRUNG IN DIE SOZIALE NETZWERKE	3
2.1	Ziele der Analyse	3
2.2	Einführung in die Grundstruktur von Netzwerken	4
2.3	Einführung in die Grundstruktur von sozialen Netzwerken	5
3	KERNFAKTOREN EINER SOZIALEN NETZWERKANALYSE	7
3.1	Gradzentralität	7
3.2	Nähe-Zentralität	8
3.3	Betweeness-Zentralität	9
3.4	Eigenvektor-Zentralität	9
3.5	Cliquen und Brücken	10
3.6	Ein typisches soziales Netzwerk	10
3.7	Kurzes Recap	12
II	DER PRAKTISCHE TEIL	
4	DER GRAPHEN GENERATOR	14
4.1	Generierung eines random Netzwerk-Plots mit vordefinierten Modellen	15
4.2	Random Graphen Generierung	17
4.3	Die Analyse des generierten Graphen	20
4.4	Die Verteilung der Zentralitäten	23
4.5	Kurzes Recap	27
5	DER VERGLEICH MIT SOZIALEN NETZWERKEN	28
5.1	Der Datensatz und die Analyse	28
5.2	Anpassung des generierten Plots	33
6	FAZIT UND AUSBLICK	36
III	ANHANG	
A	ANHANG	39
	LITERATUR	40

ABBILDUNGSVERZEICHNIS

Abbildung 2.1	Links ist Netzwerk1 als Graph dargestellt und rechts Netzwerk2	5
Abbildung 3.1	Game of Thrones social Network, Quelle: https://predictivehacks.com/social-network-analysis-of-game-of-thrones/ , Stand: 28.03.2022	11
Abbildung 4.1	Erste Versuche eines Sozialen Netzwerks, selbst erstellt	14
Abbildung 4.2	zufällig erstellte Graphen mit 25 Knoten nach den jeweiligen Methoden	16
Abbildung 4.3	Random Soziale Graphen mit den höchsten Gradzentralitäts-Knoten als Verbindung	19
Abbildung 4.4	Random soziales Netzwerk mit realistischeren Verbindungen	20
Abbildung 4.5	Verteilung der Grad-Zentralität des Graphen (b)	23
Abbildung 4.6	Random soziales Netzwerk mit realistischeren Verbindungen	25
Abbildung 5.1	Game of Thrones Graph 2.0, selbst erstellt	29
Abbildung 5.2	Game of Thrones Verteilung der Zentralitäten	30
Abbildung 5.3	Facebook Graph	31
Abbildung 5.4	Facebook Graph Distribution	32
Abbildung 5.5	Final optimierter Graph	34

TABELLENVERZEICHNIS

Tabelle 3.1	Werte GOT Graph	12
Tabelle 4.1	Werte oberer Graph	21

LISTINGS

AKRONYME

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 [11]. Auch weckt es das Interesse von Unzähligen Unternehmen, um gezielter auf das Kundenverhalten einzugehen und dadurch den Gewinn zu maximieren [10]. 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 [11][10]. 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 wir erreichen, dass der Generator Testdaten beziehungsweise Test-Netzwerke erstellt, die zum Training oder für komitative Analysen genutzt werden können. Um aber bewerten können, ob dieses Netzwerk eine gute Simulation ist, wenden wir die im ersten Teil der Arbeit vorgestellten Methoden an. 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 SOZIALE 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 themenorientierte 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 wir im Folgenden behandeln 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 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 . Diese Verbindungen werden als Kanten bezeichnet. Andernfalls werden gerichtete Verbindungen als Bögen bezeichnet. Netzwerkanten 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 [4]. 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 betrachten wir nun folgende Matrizen:

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 als Graphen abgebildet werden. Die Matrizen oben bieten sich dafür idealerweise an. Die Graphen würde in diesem Fall wie folgt aussehen:

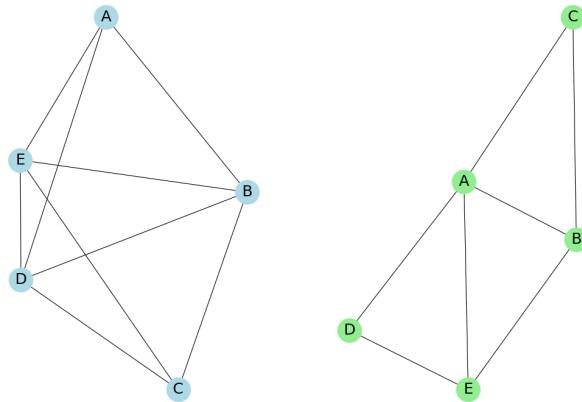


Abbildung 2.1: Links ist Netzwerk1 als Graph dargestellt und rechts Netzwerk2

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. Bekannte Programmiersprachen wie Python, besitzen bereits vordefinierte Pakete. In Python heißt dies "NetworkX". Dies ist für die Erstellung, Bearbeitung und Untersuchung von Struktur, Dynamik und Funktionen komplexer Netzwerke zuständig [2]. Hierauf werden wir zu einem späteren Zeitpunkt genauer eingehen.

2.3 EINFÜHRUNG IN DIE GRUNDSTRUKTUR VON SOZIALEN NETZWERKEN

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 Cliques, 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 [8]. 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 [5].

3

KERNAKTOREN 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 [3]. 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 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. [7]. 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 [13]. Nun folgt ein Überblick, über die wichtigsten Zentralitätsmaße und die Hauptmerkmale dieser.

3.1 GRADZENTRALITÄT

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:

$$k_i = \sum_{j=1}^N A_{ij} \quad (3.1)$$

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 Netzwerks 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 einbeziehungsweise 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 [7]. Um solche Faktoren mit berücksichtigen zu können, möchten wir einen weiteren Faktor mit in die Berechnung integrieren, nämlich die Weglänge. Diese spielt eine wichtige Rolle bei der "Nähe-Zentralität".

3.2 NÄHE-ZENTRALITÄT

Als nächstes möchten wir einen weiteren Faktor betrachten, woraus wie eine neue Formel herleiten können, nämlich die Weglänge. 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:

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

Dabei ist mit d_{ij} der kürzeste Weg zwischen i und j gemeint und mit N erneut die Anzahl an Knoten im Netzwerk [7]. Die Nähe-Zentralität ist vor allem dann sehr geeignet, wenn Prozesse über kurze Wege charakterisiert werden wollen. Betrachten wir beispielsweise den hierarchischen Aufbau eines Unternehmens. Dieses kann 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 wir nun ausschließlich anhand der Grad-Zentralität argumentieren würden, wären 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 [7].

3.3 BETWEENNESS-ZENTRALITÄT

Die Betweenness-Zentralität 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 Betweenness-Zentralität des Knotens N generiert wird. Mathematisch formuliert sieht die Formel dann wie folgt aus:

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

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:

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

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 Betweenness-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 [7]. Einem Nutzer mit hoher Betweenness-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 verbunden, 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.

3.4 EIGENVEKTOR-ZENTRALITÄT

Die Eigenvektor-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 "Hauptvektor" anhand der Adjazenzmatrix zu bestimmen. Mathematisch betrachtet ist die Eigenvektor-Zentralität die komplizierteste, der in dieser Arbeit betrachteten Zentralitäten.

Wir gehen nun von der Vorstellung aus, dass ein Akteur zentraler ist, wenn er in Beziehung

zu Akteuren steht, die selbst zentral sind. Wir können also argumentieren, dass die Zentralität eines Knotens nicht nur von 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:

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

oder umgeschrieben:

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

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

$$Ac = \lambda c \quad (3.7)$$

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 [14].

3.5 CLIQUEN UND BRÜCKEN

3.6 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 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 [9]. 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 [12]. Nun wollen wir uns jedoch auch ein typisches soziales Netzwerk anschauen. Wir müssen uns jedoch stets bewusst sein, dass es sich hierbei um einen Datensatz von einem fiktiven Fantasy Drama handelt. Genauer gesagt nehmen wir den Datensatz von Game of Thrones zu Hand und betrachten eine bereits durchgeführte SNA genauer [12]:

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

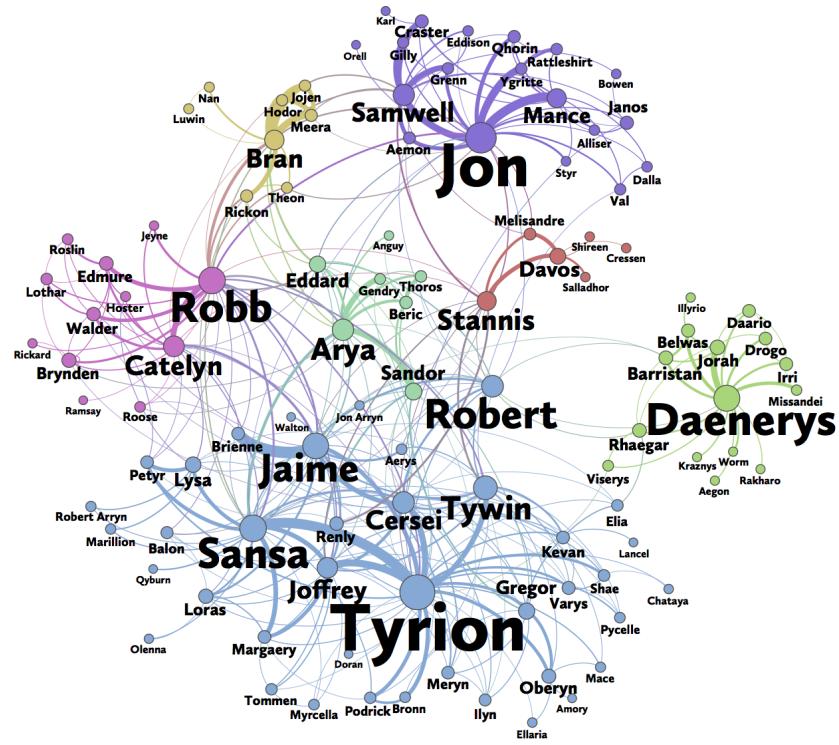


Abbildung 3.1: Game of Thrones social Network, www.socnetw.info/networks/game-of-thrones-social-network

Quelle: <https://predictivehacks.com/social-network-analysis-of-game-of-thrones/>,

Stand: 28.03.2022

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 werden wir uns aber zu einem späteren Zeitpunkt anschauen. 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 [12]. Wenn wir diese in der Abbildung 3.1 suchen, sehen wir, 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 3.1 zu analysieren. Betrachten wir die Spalten der "*Charakter*", fällt vor allem auf, dass "*Tyrion – Lannister*" in allen Spalten aufgeführt wird. Das heißt, dass dieser Knoten im Graphen sowohl zentral liegen muss, zudem kurze Abstände zu den anderen Knoten nachweisen und über diesen Knoten verlaufen zudem die häufigsten kürzesten Wege. Werfen wir nun einen Blick auf den Graphen 3.1, so sehen wir, dass der Knoten, beziehungsweise Charakter, *Tyrion* sofort auffällt. Er liegt

Tabelle 3.1: 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

zwar nicht komplett mittig im Graphen aber ist von den meisten Knoten und Kanten umgeben. Da drei der fünf wichtigsten Knoten 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 *Nhe – 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 3.1 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. Jedoch werden wir die Analyse dieses interessanten sozialen Netzwerks nicht weiterführen, sondern uns auf die Analyse des künstlich erstellen sozialen Netzwerks, zu einem späteren Zeitpunkt in dieser Arbeit, fokussieren. Auch der Frage welcher mathematische bzw. stochastische Verteilung die Zentralitäten entsprechen und warum eine solche Untersuchung sinnvoll ist, werden wir zu einem späteren Zeitpunkt nachgehen.

3.7 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 haben wir uns gemeinsam ein Beispiel für ein soziales Netzwerk angeschaut 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 wir die Theorie hinter sozialen Netzwerken und der Analyse dieser erarbeitet haben, wollen wir uns nun damit beschäftigen, wie wir typische soziale Netzwerke generieren 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 finden wir keine Cliques oder Bridgen (Brücken) in solchen Graphen, was ebenfalls dafür spricht, dass es sich um kein *typisches soziales Netzwerk* handelt.

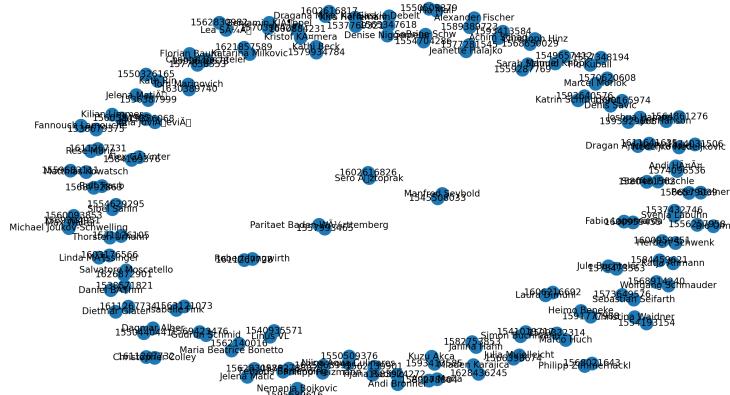


Abbildung 4.1: Erste Versuche eines Sozialen Netzwerks,
selbst erstellt

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 RANDOM NETZWERK-PLOTS MIT VORDEFINIERTEN MODELLEN

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].

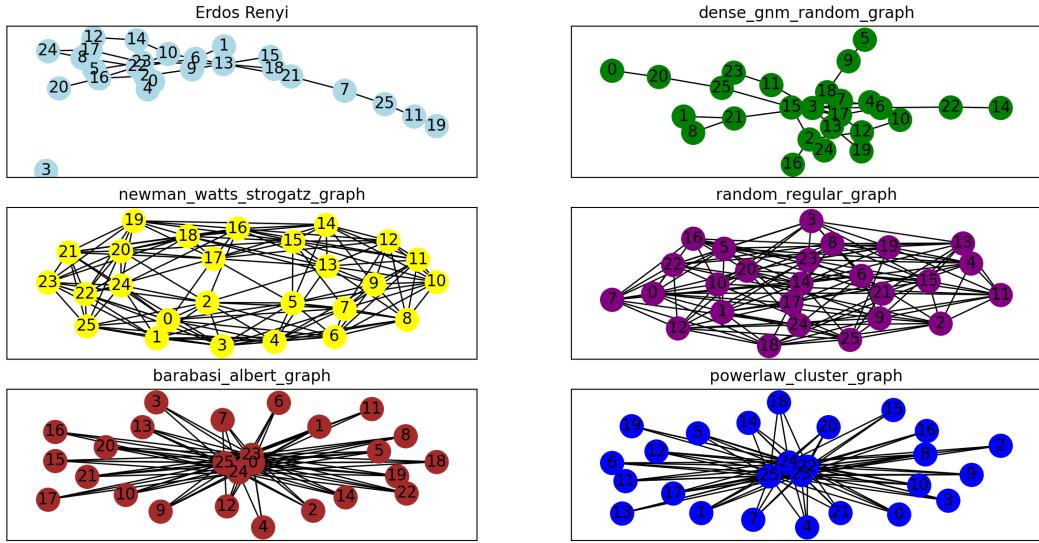


Abbildung 4.2: zufällig erstellte Graphen mit 25 Knoten nach den jeweiligen Methoden

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 erkennen wir einzelne zentrale Knoten. Diese zentralen 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 haben wir generell wenige Cliques und müssen feststellen, dass der Plot keine Cluster aufweist, weshalb dieses Modell ebenfalls nicht brauchbar ist. Bei dem *Erdos Renyi Modell* haben wir die gleiche Problematik wobei hier noch das Problem hinzukommt, dass wir einen isolierten Knoten finden. 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. Deshalb muss auch bei diesem Modell kritisch hinterfragt werden, ob es sich bei den Graphen um ein *soziales Netzwerke* handelt. Im Allgemeinen liegt die Vermutung lieg nahe, dass es bei den Netzwerken in 4.2 zu Unstimmigkeiten mit unseren Erwartungen

kommt, da wir lediglich **25** Knoten betrachten. Dies ist für ein *soziales Netzwerk* durchaus zu wenig. Deshalb liegt nahe, dass wir Anpassungen durchführen müssen.

4.2 RANDOM GRAPHEN GENERIERUNG

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

Algorithm 1 Random Adjazenzmatrix

```

1: procedure RANDOM ADJACENCY MATRIX
2:   matrix  $\leftarrow$  zufällige Matrix der Größe  $(n,n)$  zufällig gefüllt 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 benötigen wir 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  $\leftarrow$  leeres Array
3:   for i zwischen 1 und der Anzahl an Subgraphen / Matrizen do
4:     k  $\leftarrow$  zufälliger integer, der die Größe des Subgraphen definiert
5:     p  $\leftarrow$  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

```

Wir fügen also die einzelne Matrizen der Liste hinten an. Nachdem wir nun eine Liste mit vielen zufällig erzeugten Matrizen generieren konnten, fehlt uns 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 längre der Liste graphs aus nur einem Element besteht then
3:     gebe graphs zurück.
4:   dimension  $\leftarrow$  0
5:   big graph  $\leftarrow$  Graph mit Nullen befüllt
6:   for i zwischen o und der Länge von graphs do
7:     Variable a  $\leftarrow$  zufälliger integer zwischen o und Länge von graphs
8:     Variable b  $\leftarrow$  zufälliger integer zwischen o und Länge von graphs
9:     for j und k zwischen o und graphs do
10:      l  $\leftarrow$  summierte Länge von Graphs bis zur Stelle i
11:      big graph an der Stelle  $[(l+j)][(l+k)] \leftarrow \text{graph}[j][k]$ 
12:      big graph an der Stelle  $[(l+k)][(l+j)] \leftarrow \text{graph}[k][j]$ 
13:      big graph an der Stelle  $[(l+a)][(l+b+\text{graphs Länge an } [i])] \text{ modulo der Dim} \leftarrow 1$ 
14:      big graph an der Stelle  $[(l+b+\text{graphs Länge an } [i]) \text{ modulo Dim}][(l+a)] \leftarrow 1$ 
15: nun sollten wir den Knoten mit der höchsten Gradzentralität finden, um die einzelnen Subgraphen miteinander zu verbinden. Dies machen wir wie folgt
16:
17:   counter 1  $\leftarrow$  0
18:   counter 2  $\leftarrow$  0
19:   Knoten  $\leftarrow$  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  $\leftarrow$  erhöhe um 1
23:       if counter 1 größer counter 2 then
24:         counter 1  $\leftarrow$  counter 2
25:         Knoten  $\leftarrow$  i
RETURN Knoten

```

Jetzt erhalten wir einen großen Graphen, 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 Cliques besteht. Wir wollen mit diesem zusätzlich Faktor sicherstellen, dass der generierte Graph mehr Kanten besitzt und die Cluster in der Visualisierung schöne Gruppierungen aufweist. Der Cliques-Methode soll hierfür eine fixe Zahl *n* übergeben und zusätzlich sichergestellt werden, dass stetig neue Graphen generiert werden müssen, bis die Anzahl an Cliques genau der fixen Zahl *n* entspricht. Durch unsere Methoden 1, 2 und 3 erhalten wir schließlich folgenden Graphen:

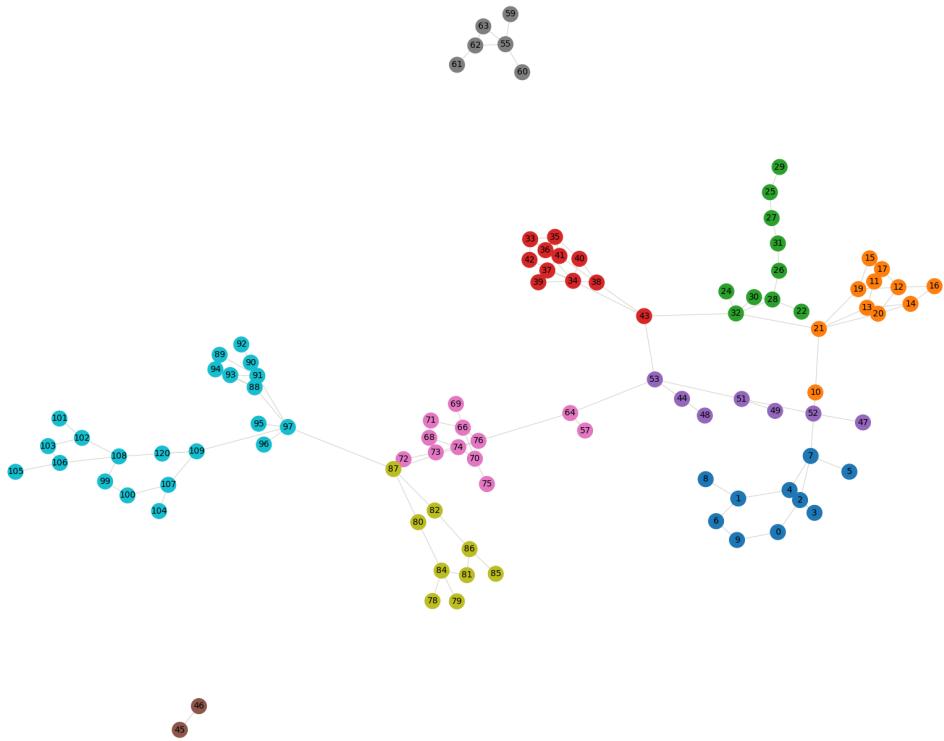


Abbildung 4.3: Random Soziale Graphen mit den höchsten Gradzentralitäts-Knoten als Verbindung

Nachdem der Plot 4.3 durchaus *Sozialen Netzwerken* ähnelt und die Werte der Berechnungen ebenfalls richtig erscheinen, müssen wir noch eine weitere Optimierung durchführen. Bei einer genaueren Betrachtung der Abbildung fällt auf, dass die Teilgraphen wenige Verbindungen 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 einer*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 müssen wir diese Eigenschaft ebenfalls in der Implementierung berücksichtigen. Das kann gewährleistet werden, indem jedem Knoten eine zufällige Wahrscheinlichkeit zugeschrieben wird, die angibt, ob eine Kante zwischen den Subgraphen existiert. Hierzu ersetzen wir den Algorithmus 3 ab Zeile 17 zu:

Algorithm 4 Verbindung Subgraphen

```

1: procedure CONNECTION SUBGRAPHS
2:   prob  $\leftarrow$  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  $\leftarrow$  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 können wir sicherstellen, dass die Subgraphen vermehrt miteinander verbunden sind und nicht nur von dem Knoten mit der höchsten Gradzentralität abhängen.

4.3 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, wollen wir ein neues generieren und eine Analyse darauf durchführen:

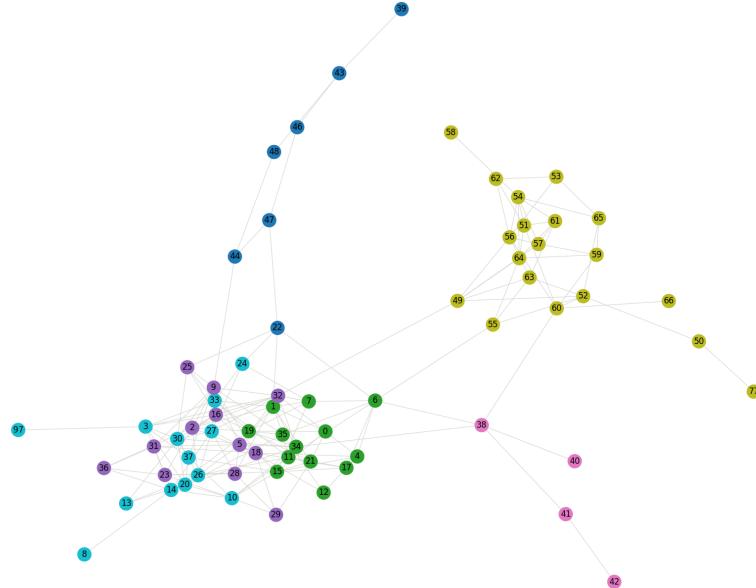


Abbildung 4.4: Random soziales Netzwerk mit realistischeren Verbindungen

Um diesen Plot zu erhalten haben wir nach wie vor primär mit Adjazenzmatrizen gearbeitet, welche erneut zufällig realisiert wurden. Statt die Knoten mit der höchsten Gradzentralität als Verbindungsglied zu wählen, wird jetzt ein beliebiger Knoten mit einer beliebigen Wahrscheinlichkeit gewählt. Dadurch entstehen Graphen, die Sozialen Netzwerken tatsächlich ähneln. Doch entsprechen die erzeugten Graphen tatsächlich näherungsweise sozialen Netzwerken? Wie kann dies bestmöglich untersucht werden? Nachdem nun viele Faktoren optimiert sind, betrachten wir den erstellen Graphen und untersuchen diesen, ob es sich um ein soziales Netzwerk handelt. Bei der objektiven Betrachtung des Plots ähnelt die Struktur auf jeden Fall der, eines sozialen Netzwerks. Doch um eine fundierte Aussagen treffen zu können, müssen die Zentralitäten erst genauer analysiert werden. Hierfür wird folgende Tabelle verwendet:

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 4.3 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 hier haben wir die Grenze rein zufällig gewählt. Bei diesem Grenzwert handelt es sich um einen guten Mittelwert. Wir wollen weder zu wenig, noch zu viele Knoten betrachten. Doch betrachten wir nun die Tabelle 4.3 genauer. Bei der Grad-Zentralität sehen wir, dass die meisten Knoten einen Wert höher als **0.1** aufweisen. Wir können einen Schritt weitergehen und stellen schnell fest, dass einige wenige Knoten eine Grad-Zentralität höher als **0.13** aufweisen. 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. Betrachten wir nun die Abbildung genauer, kann diese Behauptung teilweise bestätigt werden, denn die Knoten stechen auf jeden Fall heraus, doch befinden sie sich nicht ganz mittig im Graphen. Diese relativ hohen Werte sagen über die Knoten aus, dass es sich im realen Leben um eine sehr berühmte / bekannte Person handeln muss. Wir können beispielsweise annehmen, dass es ein Star, ein Influenzer oder eine, auf weitere Arten bekannte Person ist. Doch ebenso können wir annehmen, dass die Person lediglich viele andere Personen kennt, oder von vielen anderen Personen bekannt wird. Doch nicht nur die Grad-Zentralität spielt für uns und die Analyse in dieser Arbeit eine zentrale Rolle. Im Weiteren betrachten wir die "**Nähe-Zentralität**" doch um auch bei diesem Aspekt nicht alle 32 Werte aufzuzählen, betrachten wir im Folgenden nun Knoten, die einen Wert höher als 0.37 aufzeigen. 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 haben wir in dem ersten Teil der Arbeit gesehen, desto näher befinden sich diese Knoten zu weiteren bzw. weisen die durchschnittlich kürzesten Wege nach. Betrachten wir nach dieser Information unseren Graphen 4.4 und suchen die Knoten mit der höchsten **Nähe-Zentralität**, sehen wir direkt, dass sich diese im gleichen Bereich befinden, wie die Knoten mit der höchsten **Grad-Zentralität**. Doch bestätigt der Plot unsere Vermutung nicht eindeutig, da es teilweise nicht ideal zu erkennen ist, ob die Kanten zum Knoten verlaufen oder an diesem vorbei. Doch diese Problematik ist nicht neu für uns, sie ist bereits im ersten Teil bei der Analyse des **Game of Thrones** Graphen 3.1 aufgetreten. Wobei auch die Möglichkeit besteht, dass wir die Formeln nicht korrekt implementiert haben. Jedoch hat händisches Nachrechnen, bei durchaus kleineren Plots, die Korrektheit ergeben, weshalb wir diese Vermutung in Klammern setzen und uns eher auf die nicht eindeutigen Darstellung einigen. Doch nun kommen wir zur letzten untersuchten Zentralität, nämlich der **Betweenness-Zentralität**. Auch hier betrachten wir wieder die Knoten mit den höchsten Werten, und um nicht alle 32 Werte aufzuzählen, betrachten wir erneut nur Knoten mit einem Wert höher als 0.09. 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 anteilmäßig am öftesten über diese genannten Knoten verlaufen. Betrachten wir erneut den Graphen, sehen wir 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. In diesem Fall sehen wir sogar schön, dass die Werte gut zu unserem Plot passen und es sehr wahrscheinlich ist, dass unsere Annahme korrekt ist, und die Knoten tatsächlich am häufigsten bei allen kürzesten Wegen durchlaufen werden. Weitere Zentralitätswerte müssen wir nicht betrachten. Zum einen gäbe es hier noch die **Eigenvektor-Zentralität**, doch würde diese nur unsere Feststellungen bestätigen. Jetzt haben wir alle Kriterien überprüft und erfolgreich festgestellt, dass dieser Graph einem **sozialen Netzwerk** ähnelt. Doch um auszuschließen, dass wir hier rein zufällig ein solches erhalten haben, möchten wir uns noch ein weiteres Kriterium überlegen und anschließend untersuchen.

4.4 DIE VERTEILUNG DER ZENTRALITÄTEN

Nachdem wir im vorherigen Kapitel eine **soziale Netzwerkanalyse** durchgeführt haben und ein gutes soziales Netzwerk künstlich generiert haben, möchten wir uns im Folgenden die Verteilung der Zentralitätswerte anschauen. Im Laufe der Arbeit ist aufgefallen, dass sich die Werte der Zentralitäten oftmals in ähnlichen Bereichen befinden. An dieser Stelle können wir uns die Frage stellen, wie diese Werte verteilt sind. Vielleicht existieren Zusammenhänge zwischen unseren generierten sozialen Netzwerken und sozialen Netzwerken im Allgemeinen. Das heißt, im Konkreten, wollen wir der Frage nachgehen, ob alle Zentralitätswerte sozialer Netzwerke ähnliche Verteilungen nachweisen. Würde sich unsere Vermutung diesbezüglich bestätigen, können wir andere soziale Netzwerke anhand dieses Kriterium untersuchen und interessante Vermutungen aufstellen. Zunächst aber implementieren wir die Methode, die die Verteilung der Zentralitäten untersucht anhand der **Gradzentralität**. Da der Graph 4.4 ein zufällig, einmalig erzeugter Graph ist, werden wir nicht einen Identischen Graphen erzeugen können, um die Verteilung der Zentralitäten zu betrachten. Was sich jedoch nicht negativ auf die weiteren Untersuchungen auswirken sollte. Den ersten Graphen und die zugehörige Verteilung der Gradzentralität bildet sich folgendermaßen ab:

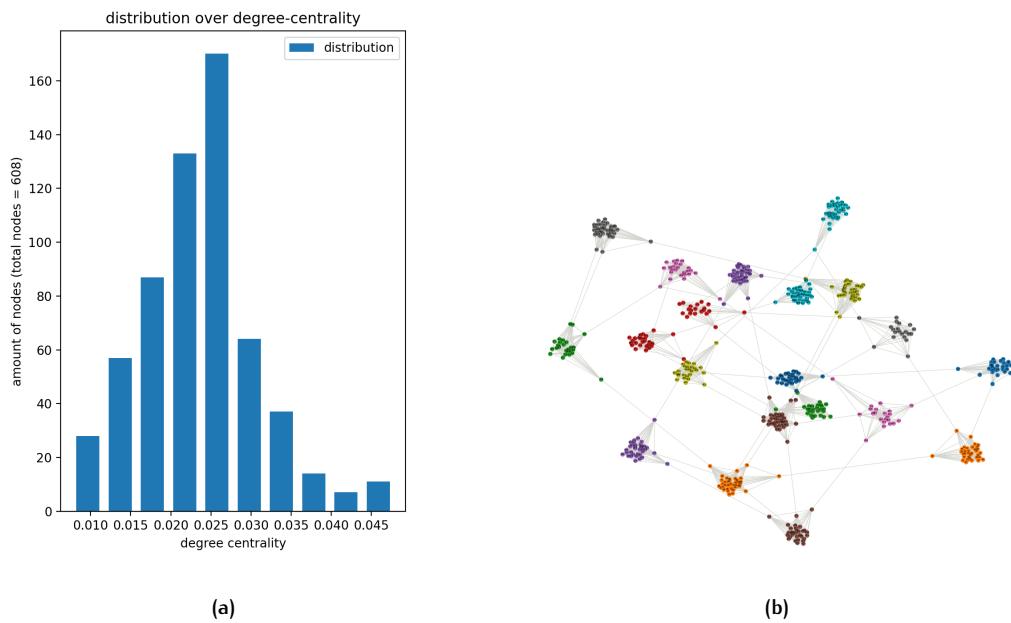


Abbildung 4.5: Verteilung der Grad-Zentralität des Graphen (b)

Jedoch ist zu erwähnen, dass wir keine perfekte Gauß-Verteilung generieren, sondern eine etwas nach links verschobene Verteilung sehen. An was dies liegen kann, werden wir uns später anschauen und versuchen das Phänomen zu analysieren. Nun wollen wir untersuchen, ob sich die Eigenschaft, der gleichmäßig verteilten Zentralitäten für die **Nähe-** und **Betweenness-Zentralität** bestätigt. Um zusätzlich zu beweisen, dass es sich bei der gleichmäßigen Verteilung der Werte nicht um einen Zufall handelt, generieren wir einen neuen sozialem Graphen und

untersuchen die Verteilung der **Grad-, Nähe-, Betweenness- und Eigenvektor-Zentralität**. Hierbei stellen wir uns vor allem die Frage, entspricht die Verteilung einer tatsächlichen Normalverteilung und falls ja, warum ist dies der Fall. Ansonsten stellen wir uns die Frage warum es keiner Normalverteilung entspricht und ob es möglich wäre, den Graphen zu verändern um eine zu erzielen.

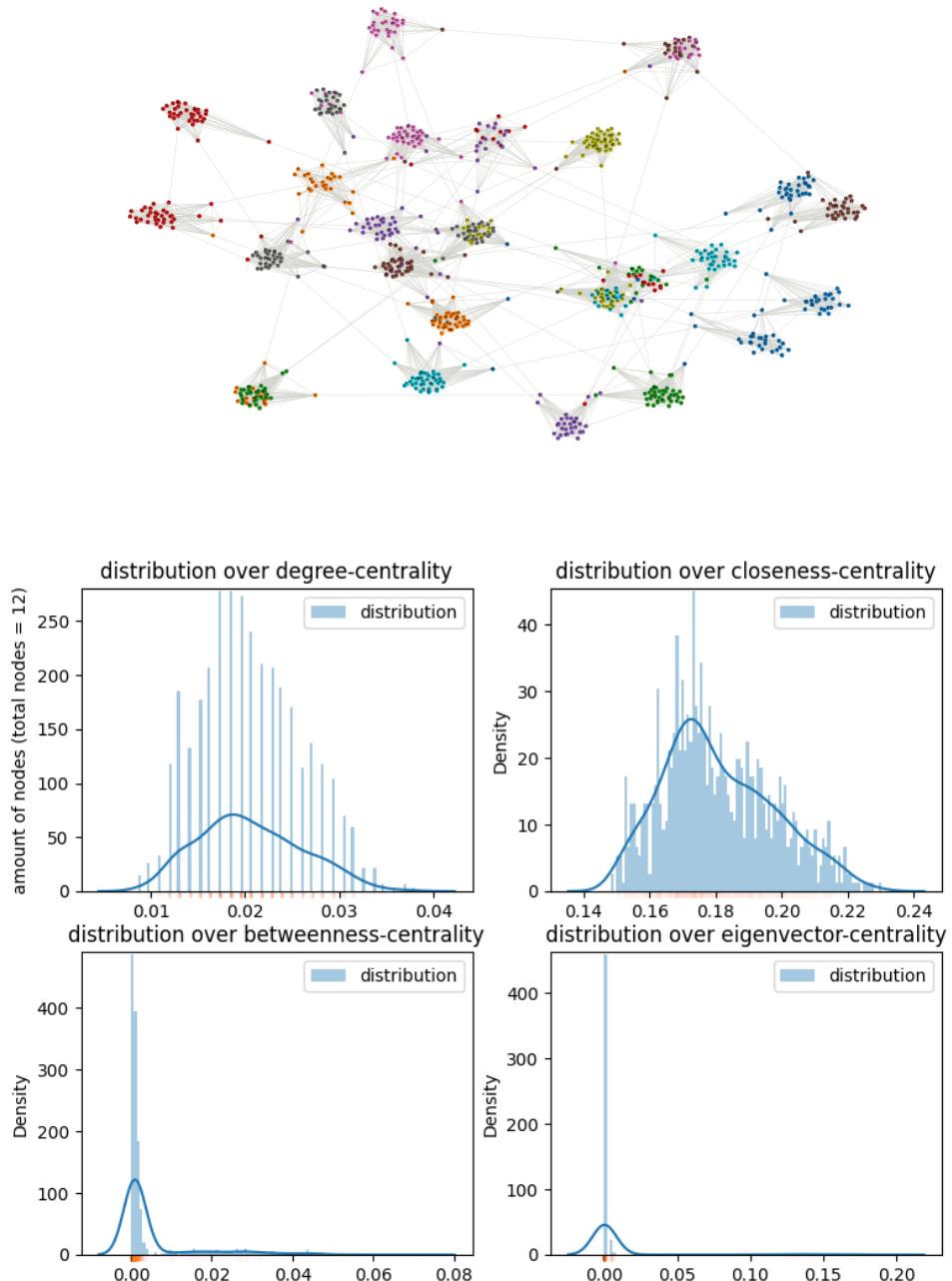


Abbildung 4.6: Random soziales Netzwerk mit realistischeren Verbindungen

In der Abbildung 4.6 sehen wir nun die Verteilungen der Zentralitäten des, sich darunter befindenden sozialen Netzwerks. Die Tabelle mit den Zentralitäts-Werten des Netzwerks

befindet sich in A. Oben links sehen wir die Verteilung der **Grad-Zentralität**, welche wie bereits oben festgestellt, nicht perfekt normalverteilt ist. Vor allem auffällig ist, dass der Balken bei **0.012** sehr schnell vergleichsweise stark sinkt. Über **50** Knoten weisen diesen Wert nach, danach geht der Balken nochmals zurück, denn nur noch etwas über **25** Knoten haben eine Zentralität von etwas unter **0.03**. Wir hätten aber erwartet, 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 lobend zu erwähnen, dass wir genau **einen Peak** erreichen, wie wir auch erwartet haben. Zudem sind alle Balken vor dem Peak kontinuierlich aufsteigend und nach dem Peak kontinuierlich absteigend. Doch eine Unstimmigkeit sticht hier heraus, bei dem Zentralitätswert von **0.0357** und etwas unter **25** Knoten, die diesen Wert aufweisen. Fraglich ist hierbei, warum der Balken erneut gewachsen ist. Denn im Regelfall sollten wir maximal ein bis drei Knoten finden, die diesen Wert aufweisen. Doch im Allgemeinen weist der Plot genau die Eigenschaft nach, die wir auch erwartete haben, nämlich dass die **Grad-Zentralität** annähernd normalverteilt ist. Wenn wir nun zu dem Balkendiagramm der **Nähe-Zentralität bzw. closeness-centrality** blicken, sehen wir einen ähnlichen Verlauf. Wir entdecken einen bereits erwarteten 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. Wir haben viele Generierungen durchgeführt und in A befinden sich ebenso weitere Balkendiagramme und Soziale Netzwerke. Aussagen die wir auf jeden Fall sicher treffen können ist, dass falls es zu Unstimmigkeiten kommt diese stets an anderen Stellen auftreten und nicht immer den letzten Balken betreffen. Diese Aussage können wir jedoch nur für die Verteilung der **Grad- und Nähe-Zentralitäten** treffen. Jedoch könne wir ebenso annnehmen, dass je größer der Graph ist, umso eher sind diese Zentralitäten normalverteilt. Was daran liegt, dass wir mehr Knoten haben und diese irgendwann eine Regelmäßigkeit aufzeigen. 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 schauen wir uns auch noch die zwei unteren Balkendiagramme an, wird unsere Behauptung verworfen. Bei der **Betweenness- und Eigenvektor-Zentralität** wird unsere angenommene Regelmäßigkeit nicht bestätigt. Zum einen weisen die Balken wenige unterschiedliche Werte auf, die teilweise kaum zu unterscheiden sind. Was hingegen auffällt, sind die Ausschläge der hintersten Balken. Was zunächst verwunderlich erscheint, ist mit einer simplen Erklärung begründet. Die **Closeness-Zentralität** gibt bekanntlich an, wie oft ein Knoten anteilmäßig 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** haben wir zwar die gleiche Beobachtung gemacht doch sagt diese hier etwas anderes aus. Diese Zentralität gibt uns eine Einschätzung der Wichtigkeit des Knotens, im Bezug auf seine Nachbarn an, was bezogen auf unser Balkendiagramm heißt, dass viele Knoten in unserem Graphen wichtig sind mit Einbeziehung der Nachbarn. Wobei wir auch vermuten können, dass dies mit der hohen Anzahl an Konten mit höher **Betweenness-Zentralität** zusammenhängt. Die komplette Analyse des sozialen Netzwerks in 4.6 befindet sich in A, da wir uns bereits eine ausführliche Analyse durchgeführt hatten. Danach zu Urteilen handelt es sich bei dem Netzwerk um ein typisches

soziales Netzwerk. Doch wie hängt jetzt unsere Verteilung der Zentralitäten mit der Verteilung der Zentralitäten von anderen sozialen Netzwerken ab?

4.5 KURZES RECAP

Nachdem wir uns nun angeschaut haben, wie wir soziale Netzwerke mithilfe Python Funktionen generieren können konnten wir auch gleichzeitig die Probleme dieser sehen. Anschließend haben wir unseren Code fortlaufend optimiert und ein soziales Netzwerk erstellt. Für dieses haben wir danach eine soziale Netzwerk Analyse erstellt und festgestellt, dass es die Anforderungen an ein soziales Netzwerk erfüllt. Danach ist uns aufgefallen, dass die Zentralitäten regelmäßig sind und konnten die Normalverteilung nachweisen. Doch müssen wir uns im Folgenden die Frage stellen, wie die Verteilung der Zentralitäten bei anderen, bereits analysierten Netzwerken ist. Genau das machen wir im folgenden Kapitel.

5

DER VERGLEICH MIT SOZIALEN NETZWERKEN

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 vielleicht sogar bessere Graphen zu erhalten, die diesen 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 Plots, über Sozialen Netzwerken, 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 [12] von unserem **Game of Thrones** Plot 3.1. Da wir bereit die Analyse der **Zentralitäten** und die generelle visuelle Analyse des Graphen durchgeführt haben, reicht uns nun lediglich die Verteilung der Zentralitäten durchzuführen. Die Tabelle mit den Werten der Zentralitätsberechnungen befinden sich erneut in A. Nachdem wir den Datensatz als **CSV** Datei in Python eingelesen und zunächst den Graphen folgendermaßen konstruiert:



Abbildung 5.1: Game of Thrones Graph 2.0,
selbst erstellt

Dieser Plot ist beabsichtigt klein, da wir ihn lediglich zur Argumentation für die Verteilung der Zentralitäten benötigen und daher die Form des Graphen nur von Bedeutung für uns ist. Zudem ist zu vermerken, dass der eigentliche Datensatz gewichtet ist, und unser Graph daher bereits schaßvisuell nicht dem Graphen aus 3.1 ähnelt. Jedoch ist es sinnvoll die Gewichte außen vor zu lassen, da wir in dieser Arbeit ungewichtete Graphen nachbilden. Nachdem wir die Daten des Graphen 5.1 eingelesen, die Zentralitäten berechnet haben und anschließend den Balkengraphen erstellt haben, wurde folgender Plot generiert:

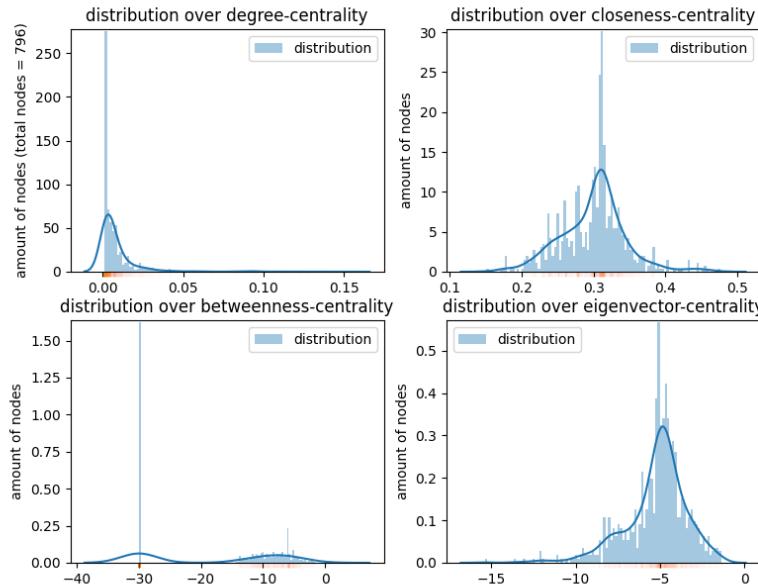


Abbildung 5.2: Game of Thrones Verteilung der Zentralitäten

Auf den ersten Blick können wir bereits feststellen, dass wir andere Ergebnisse erwartet haben. Einzig die Verteilung der **Nähe-Zentralität** ähnelt der erwarteten Normalverteilung. Die **Betweenness-** und **Eigenvektor-Zentralität** hingegen ähneln zwar nicht exakt dem, was wir in 4.6 herausgefunden haben aber ziehen auf jeden Fall Parallelen. Denn beide haben den Ausschlag im letzten Balken, was wir schon im vorherigen Kapitel damit begründet haben, dass es die Folge davon ist, wenn viele kürzeste Wege stets über die gleichen Knoten verlaufen, wir also keine Alternativen im Graphen haben. Die **Grad-Zentralität** hingegen darf uns verwundern. Sie ähnelt zum einen stark der Verteilung der **Eigenvektor-Zentralität** aber keinesfalls der annähernden normalverteilt aus 4.6. Der Ausschlag des letzten Balken ist hingegen schnell erklärt. Wir haben viele Konten, in dem Fall Charaktere, die alle gleich wichtig für den Graphen sind. Diese Konten sind also mit vielen anderen Knoten verbunden, werden daher von vielen anderen Charakteren bekannt 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, 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 geführt hätte. Doch wir haben den Datensatz aber ungewichtet betrachtet, um ihn besser mit unseren generierten Graphen zu vergleichen, welche ungewichtet sind. Dies kann auf jeden Fall ein plausibler Grund für Unstimmigkeiten sein. Zudem haben wir die Anzahl der geplotteten Balken stark erhöht und so fallen Unstimmigkeiten auch deutlich stärker auf. Doch wollen wir unsere Theorie, dass Zentralitäten normalverteilt sind, nicht verwerfen und wollen uns ein bis zwei weitere Datensätze anschauen. Als nächstes betrachten wir einen Datensatz, der aus aus "Kreisen"(oder "Freundeslisten") von Facebook besteht. Die Daten wurden anschließend

anonymisiert. So können wir mit dem Datensatz also feststellen, ob zwei Nutzer die gleiche politische Zugehörigkeit haben, aber nicht, was ihre individuelle politische Zugehörigkeit bedeutet [6]. Nachdem wir die Daten wieder in eine .CSV Datei umgewandelt und anschließend geplottet haben, konnten wir folgenden Graphen generieren:



Abbildung 5.3: Facebook Graph

Der Graph ähnelt auf den ersten Blick durchaus dem erstellten Plot 4.4. Sofort fällt aber auf, dass dieser Graph aus deutlich mehr Knoten besteht, zudem weniger Subgraphen aber dennoch im Grunde eine ähnliche Struktur aufweist. Die Berechnungen der Zentralitäten befinden sich auf Github [20]. Nun interessiert uns jedoch, wie diese Zentralitäten verteilt sind und ob dieser Graph die erwarteten Verteilungen erfüllen kann. Nachdem wir den Datensatz durch den Code laufen lassen haben, konnten wir folgenden Plot generieren:

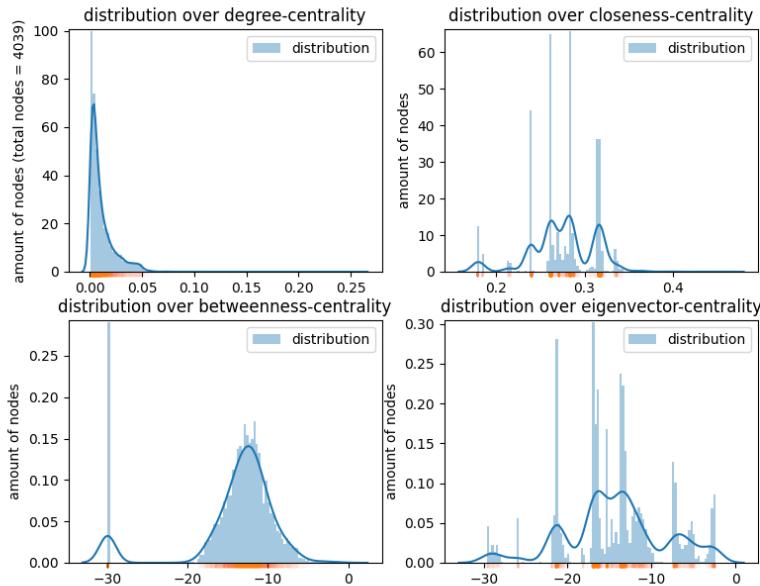


Abbildung 5.4: Facebook Graph Distribution

Sofort fällt auf, dass wir bei keiner Zentralität eine Normalverteilung sehen. Die **Grad-Zentralität** fällt uns aber direkt auf, denn es handelt sich hier um eine Exponentialverteilung. Die anderen Balkendiagramme der **Betweenness- und Eigenvektor-Zentralität** ähneln jedoch stark 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 ernüchternd scheinen und vor allem das Balkendiagramm der **Nähe-Zentralität** sehr eigen. Wir erkennen eine starke Fluktuation der Balken und daher keine schöne Verteilung, die wir aus der Mathematik als Vergleich anbringen können. Visuell fällt aber visuell auf, dass der Graph 5.4 verglichen mit dem Plot des Graphen 2 durchaus Parallelen aufweist. Wir sehen deutliche Ansammlungen von Knoten die auch gut als Teilgraphen bezeichnet werden können. Zwischen den Teilgraphen erkennen wir, so wie bei 4.4 einige Kanten, die die Teilgraphen untereinander verbinden. Natürlich weist der obige Graph deutlich mehr Kanten und Knoten auf. Unsere Graphen haben im Schnitt um die 950 Knoten und 8700 Kanten, daher also circa neun mal so viele Kanten wie Knoten. Auch haben wir im Schnitt um die 10100 Cliques, welche maximal acht Knoten groß sind. Bei dem Facebook Graphen 5.3 hingegen sprechen wir von 4093 Knoten und 88234 Kanten. Das heißt circa einundzwanzig mal so viele Kanten wie Knoten. Doch als wir unsere Kanten und Knoten im Code erhöht haben, um ebenfalls die selbe Relation zu erhalten, waren alle vier untersuchten Zentralitäten annährend Normalverteilt. Was zum einen daran liegt, dass wir letztendlich Graphen wie 5.1 erhalten haben, die noch viel dichter besetzt waren. Wir können festhalten, dass sich die **Betweenness- und Eigenvektor-Zentralität** bei allen untersuchten Datensätzen stark zu unseren Verteilungen des künstlich generierten sozialen Netzwerk ähneln. Auch die **Nähe-Zentralität** weist stets ähnliche Verteilungen auf, was die Korrektheit des Graphen 4.4 bestätigt. Doch wundert uns nach wie vor die Verteilung der **Gradzentralität**.

Daher ist es ratsam, unseren Code und damit verbundene Plot so anzupassen, dass es dem Plot 5.4 ähnelt. Anschließend können wir die Verteilungen der Zentralitäten betrachten und dadurch eine genauere Analyse garantieren.

5.2 ANPASSUNG DES GENERIERTEN PLOTS

Nachdem uns die Verteilungen durchaus gewundert haben, wollen wir den nächsten Plot weitestgehend versuchen an 5.4 anzupassen. An dieser Stelle muss durchaus betont werden, dass es sich bei unseren generierten Plot keinesfalls um untypische oder falsche soziale Netzwerke handelt. In diesem Abschnitt wollen wir lediglich eine bessere Vergleichsbasis herstellen. Dies funktioniert sehr einfach, indem wir zum einen die Anzahl an Cluster auf **sieben** anpassen und die Anzahl der Knoten pro Cluster erhöhen. Gleichzeitig aber die jeweiligen Größen deutlich mehr variieren lassen und vor allem die Kanten-Menge, also Anzahl an Verbindungen, deutlich erhöhen. Schließlich erhalten wir folgende Graphen:

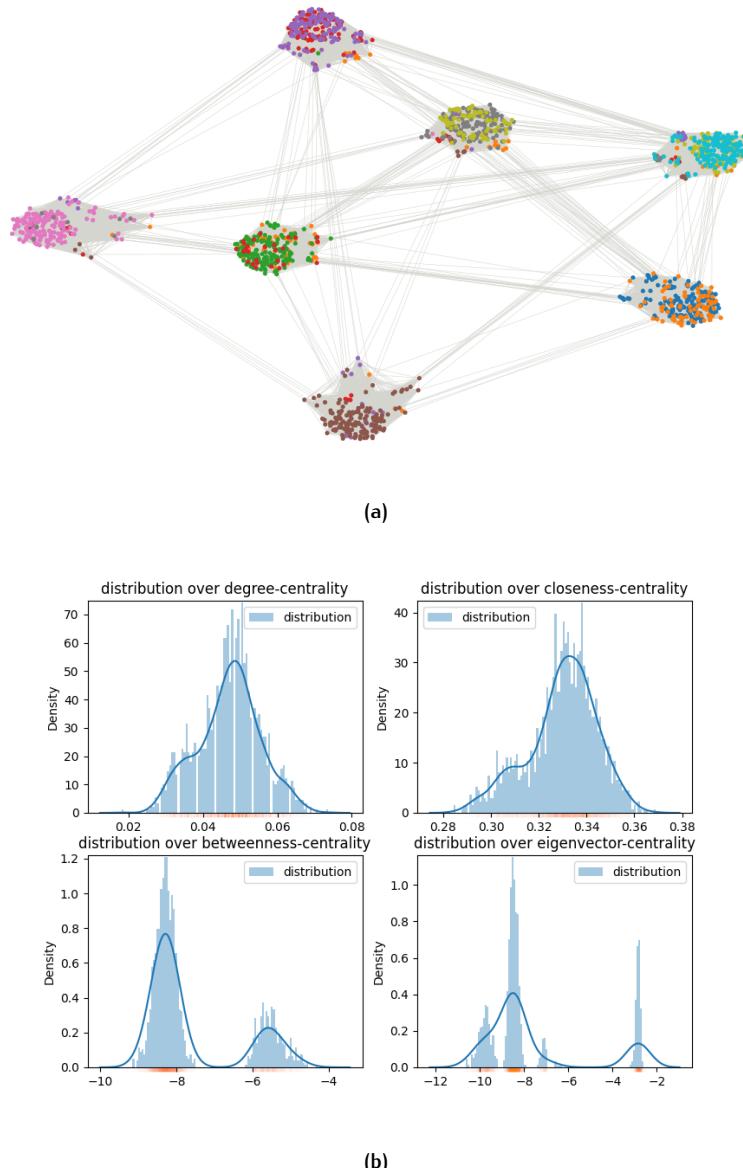


Abbildung 5.5: Final optimierter Graph

Natürlich sehen wir direkt, ohne die Werte genauer analysiert zu haben, dass wir keinen zu 5.4 identischen Graphen erzeugen konnten. Dies liegt an mehreren Faktoren. Zum einen sind wird unser großer Graph aus sieben Teilgraphen zusammengefügt, was dazu führt, dass kaum bis keine Knoten erhalten, die zwei und weniger Kanten verbinden und sich zwischen den Teilgraphen befinden. Allgemein sind unsere Graphen, auch wenn wir die Varianz der Graph-Größen best möglichst so optimal wie möglich gestalten wollten, auf den ersten Blick ähnlich groß. Jedoch haben wir bei der Verteilung der Zwischen- und Eigenvektor

Zentralität eine absolute Verbesserung erzielen können, indem wir die Zentralitäten, bevor sie im Seaborn Graphen geplottet werden, logarithmiert haben. Dies ermöglicht es uns, die Verteilung auseinander zu zerren, da wir sehen konnten, dass sich die Werte stets um 0.0 verteilt haben. Zwar haben wir nun erneut bei 5.5 und 5.4 keine identischen Verteilungen erzielen können, doch liegt die Vermutung nahe, dass wir eine Exponential-Verteilung der Gradzentralität erzielen werden. Tatsächlich hat dies aber keine gravierende Auswirkung, denn mathematisch betrachtet gilt, 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 uns diese Optimierung einige Informationen gewinnen lassen. Wir schaffen es mit wenigen Anpassungen des Codes annähernd gut vergleichbare Verteilungen zu generieren. Um den idealen Vergleich herzustellen, müssten wir jedoch große Änderungen am Code vornehmen, was jedoch nicht mehr im Umfang dieser Arbeit liegt. Zumal wir bei 5.3 nicht den Inbegriff an sozialem Netzwerk vorliegen haben, was uns bereits zum Fazit führt.

noch mehr
zu Cliques
und Brücken

6

FAZIT UND AUSBLICK

Nachdem wir uns in dieser Arbeit ausführlich mit der Generierung und Analyse sozialer Netzwerk beschäftigt haben, wollen wir 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. Wir konnten in dieser Arbeit zeigen, dass die Betrachtung der Cliques und Brücken bei der visuellen Interpretation durchaus ausreichen. Bei der Analyse der Daten wurde deutlich, dass die ausschließliche Betrachtung der einzelnen Zentralitäten bei großen Datensätzen nicht optimal ist und viel Nacharbeit in Anspruch nimmt, um die wichtigsten Knoten herauszufiltern und keine hunderttausende Werte zu vergleichen. Vielmehr ist es aussagekräftig in solchen Fällen die Verteilungen zu betrachten. Doch gleichzeitig müssen wir 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 typische soziale Netzwerke. Bei dem bloßen Vergleich von Zentralitäten sind die visuellen Differenzen oder Unstimmigkeiten zunächst nicht von Bedeutung, doch spätestens bei der Analyse der Verteilungen wird diese bedeutsam. Wir konnten jedoch in dieser Arbeit sehen, dass wir es durch Optimierungen im Code schaffen, leicht visuell ähnliche Graphen zu erzeugen. Sobald die Graphen gleiche Grundstrukturen aufweisen, bestehen starke Gleichheiten in der Verteilung der Zentralitäten. Diese Optimierung kann unendlich lange fortgeführt werden, doch erreichen wir so lediglich Kopien von existierenden Netzwerken und wertschätzen die Individualität dieser nicht. Diese Arbeit lässt einige Punkte offen, die durchaus weiter optimiert werden können. Beispielsweise die generierten Plots noch besser an existierende Graphen anpassen, um die Verteilung bestmöglich nachzustellen. Zudem größere Datensätze untersuchen und vergleichen, was sich bei Millionen von Knoten an den Zentralitäten, der Verteilungen dieser, Cliques und Brücken verändert. Vor allem zu untersuchen, ob womöglich Regelmäßigkeiten auftreten. Oder ebenfalls interessant, ob die Berechnungen der Zentralitäten bereits optimierte Algorithmen sind und ob nicht möglicherweise doch Optimierungspotenzial besteht. Ebenfalls interessant statt der Analyse eine Interpretation der Graphen durchzuführen. Bei gegebenen Graphen, ohne Vorwissen über die Datensätze, Vermutungen aufzustellen und diese auszuführen. Dies und vieles mehr sind weitere Eigenschaften, die untersucht werden können. Was uns wieder zeigt, wie vielfältig die soziale Netzwerkanalyse ist und warum sie zahlreiche Wissenschaftler*innen beschäftigt. Dank ihr konnten alleine im geschichtlichen Aspekt des Menschen viele Fragen geklärt oder zumindest Vermutungen aufgestellt und belegt werden. Doch nicht nur in der Vergangenheit spielt die Analyse eine große Rolle, im Bereich des Sozial Media Booms ist sie aktuell und wird auch garantiert in der Zukunft vieles beeinflussen. Seien es Unternehmen, die dadurch bekannte *Influencer* finden können oder App-Entwickler, die Korrelationen zwischen Nutzer*innen auf diese Weise darstellen und Apps weiter anpassen, um ihre Nutzeranzahl zu erhöhen. Schließlich können wir diese Arbeit damit beenden, dass es unglaublich vielzählige Methoden zu Analyse von Netzwerken gibt und jede Vor- und Nachteile aufweist. Welche die geeignete ist, lässt sich nicht in einem Satz formulieren. Es kommt auf Anzahlen von Kanten und Knoten an, aber auch auf die zu untersuchende Thematik.

Optimaler weise ist es stets ratsam mehrere Methoden zu verwenden, denn diese unterscheiden sich ebenfalls in ihrer Aussagestärke.

Teil III
ANHANG

A | ANHANG

LITERATUR

- [1] NetworkX Developers. *Graph generators*. 2014-2022. URL: <https://networkx.org/documentation/stable/reference/generators.html> (besucht am 28.03.2022).
- [2] NetworkX Developers. *NetworkX, Network Analysis in Python*. 2014-2022. URL: <https://networkx.org/> (besucht am 28.03.2022).
- [3] 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>.
- [4] Riddle M. Hanneman R. *Introduction to Social Network Methods (Hanneman)*. University of California, Riverside, 2019. URL: <https://math.libretexts.org/@go/page/7645>.
- [5] Charles Kadushin. "Introduction to Social Network Theory". In: (Jan. 2004).
- [6] By Jure Leskovec. *Social circles: Facebook*. 2012. URL: <https://snap.stanford.edu/data/ego-Facebook.html> (besucht am 28.03.2022).
- [7] Elbert E N Macau. *A mathematical modeling approach from nonlinear dynamics to complex systems*. Springer, 2019. URL: <https://www.worldcat.org/title/mathematical-modeling-approach-from-nonlinear-dynamics-to-complex-systems/oclc/1117866920>.
- [8] 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](https://doi.org/10.1016/S0378-8733(02)00016-3).
- [9] 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).
- [10] 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).
- [11] Ioannis Panges. *Social Network Analysis. An Introduction*. GRIN Verlag, 2016. URL: <https://www.grin.com/document/371489>.
- [12] 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).
- [13] Francisco Rodrigues. "Network Centrality: An Introduction". In: März 2018. ISBN: 978-3-319-78511-0. DOI: [10.1007/978-3-319-78512-7_10](https://doi.org/10.1007/978-3-319-78512-7_10).
- [14] Britta Ruhnau. "Eigenvector-centrality — a node-centrality?" In: *Social Networks* 22 (Okt. 2000), S. 357–365. DOI: [10.1016/S0378-8733\(00\)00031-9](https://doi.org/10.1016/S0378-8733(00)00031-9).
- [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/bachelor-thesis-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