Angewandte Computer- und Biowissenschaften

HOCHSCHULE MITTWEIDA
University of Applied Sciences

MASTERARBEIT

Herr
Sandro Martens

Überführung von Online-Kommentaren in eine Netzwerktopologie

2021

Fakultät **Angewandte Computer- und Biowissenschaften**

MASTERARBEIT

Überführung von Online-Kommentaren in eine Netzwerktopologie

Autor:
Sandro Martens

Studiengang:
Cybercrime/Cybersecurity

Seminargruppe:
CY19wC-M

Matrikelnummer:
41026

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Dr. rer. nat. Michael Spranger

Mittweida, November 2021

Bibliografische Angaben

Martens, Sandro: Überführung von Online-Kommentaren in eine Netzwerktopologie, 85 Seiten, 39 Abbildungen, Hochschule Mittweida, University of Applied Sciences, Fakultät Angewandte Computer- und Biowissenschaften

Masterarbeit, 2021

Satz: LaTeX

Referat

In dieser Arbeit werden drei Modelle entworfen und verglichen, mit welchen *Meinungsführer* in einem Twitter-Netzwerk erkannt werden können. Dazu wird ein Datensatz mit 600.000 Tweets von 100.000 Twitter-Nutzern von April bis Juni 2021 ausgewertet. Zur Bestimmung des Einflusses eines Nutzers werden sowohl topologische Informationen des Netzwerkes als auch Reaktionen auf einzelne Tweets einbezogen. Anschließend werden Korrelationen zwischen dem Grad der Meinungsführerschaft und der Toxizität der Tweets untersucht. Dafür wurde eine Recherche zu Software zur Analyse von Graphen durchgeführt und Neo4j als passendes Werkzeug ausgewählt. Es konnte gezeigt werden, dass der ArticleRank als Zentralitätsalgorithmus geeignet ist, Meinungsführer zu erkennen. Meinungsführer sind weniger toxisch als andere Nutzer, allerdings ist dies nur ein schwacher Indikator. Durch die Modellierung der Häufigkeit, wie oft Nutzer interagieren, können verschiedene Fragen beantwortet werden. Durch diesen Algorithmus können Konzepte der Kommunikationswissenschaft in Bezug auf Meinungsführer in sozialen Netzwerken nachgeweisen werden.

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

IV. Abkürzungsverzeichnis

APOC ..... Awesome Procedures On Cypher, Seite 34

BSI ..... Bundesamt für Sicherheit in der Informationstechnik, Seite 11

GDS ..... Graph Data Science, Seite 19

GDSL ..... Graph Data Science Library, Seite 34

Neo4j ..... Network Exploration and Optimization 4 Java, Seite 5

SAP ..... Sicherung, Analyse, Präsentation, Seite 11

SNA ..... Soziale Netzwerkanalyse, Seite 11

Teil I

Grundlagen und Recherche

1 Einleitung

> Das Gute am Internet ist, dass jeder seine Meinung kundtun kann. Das Schlechte ist, dass es auch jeder tut.
>
> *(Marc-Uwe Kling,* Die Känguru-Chroniken*)*

Soziale Netzwerke stellen einen wichtigen Grundbaustein der Kommunikation vieler Menschen dar. Knapp 50% der deutschen Bürger benutzen soziale Netzwerke (Statista, Statista Digital Market Outlook 2019). Diese weite Verbreitung und die Geschwindigkeit und Einfachheit, mit welcher im Internet kommuniziert werden kann, führt dazu, dass viele Ideen und Themen auf diesen Plattformen diskutiert werden. Dadurch wird es für viele Interessengruppen möglich, über diese Plattformen Einfluss auf die öffentliche Meinung sowie politische Entscheidungsbildung zu nehmen (Fernquist u. a. 2020). Dies können politische Parteien sein, aber auch nichtstaatliche Akteure wie Nichtregierungsorganisationen (NGO) oder Trolle. Diese Online-Diskussionen sind häufig polarisierend bis hasserfüllt. Die Erkennung und Vorhersage von Hasskommentaren ist eine der großen aktuellen Herausforderungen von Webseiten-Betreibern. Dies liegt einerseits an gesetzlichen Grundlagen wie dem Netzwerkdurchsetzungsgesetz, welches Betreiber von Webseiten verpflichtet, Hassnachrichten auf ihren Plattformen zu löschen. Andererseits dient dies auch zur Moderation von Gruppen und Seiten, um den Administrator Möglichkeiten zu bieten, auf kontroverse und bösartige Kommentare zu reagieren.

Facebook löschte von April bis Juni 2021 31,5 Millionen Beiträge wegen Hate Speech-Verstößen (Facebook [o. D.]). Dies ist 12,6 mal so viel wie noch von Januar bis März 2018. Twitter löschte im Zeitraum von Juli bis Dezember 2020 4,5 Millionen Beiträge. Den mit Abstand größten Anteil daran hatten *Abuse/Harassment* und *Hateful conduct* (Twitter [o. D.]). Bereits 2019 wurden bei Twitter über 50 Prozent der Hasskommentare proaktiv, also ohne Meldung eines Nutzers oder eine Meldestelle, gelöscht (Twitter 2019). 77 Prozent der (deutschen) hasserfüllten Beitrage sind laut BKA rechts motiviert (Abb. 1.1)(Brandt 2019). In Deutschland gingen von Januar bis Juni 2021 774.888 Beschwerden von Nutzern sowie 58.514 Beschwerden von Beschwerdestellen ein. Dabei wurden in 10 Prozent beziehungsweise sieben Prozent Maßnahmen ergriffen (Twitter 2021). Laut einer Studie von YouGov und Statista sind insbesondere Politiker und Aktivisten Opfer von Hass im Netz. Besonders für AfD-Wähler ist dies legitim (Inhoffen 2019).

Das Konzept Meinungsführer existiert seit etwa 1940 (Dressler u. a. 2009, S. 26). Lazarsfeld, Berelson und Gaudet führten Studien durch, wie Bürger der USA bei Präsidentschaftswahlen entscheiden. Dabei stellten sie fest, dass sich viele Menschen von

Abbildung 1.1: Verteilung von Onlinehass 2018 laut BKA.

Bekannten mehr beeinflussen lassen als von Massenmedien (Taddicken 2015). Heutzutage sind Meinungsführer insbesondere für Marketing relevant (Schach u. a. 2018). Firmen benutzen Influencer, um deren Reputation und Bekanntheit für sich zu nutzen. Dies kann negative Auswirkungen haben, wenn die Grenzen zwischen Wissenschaft beziehungsweise Journalismus und Werbung verschwimmen (Nicholas 2021).

Dieser Arbeit stand ein Datensatz mit circa 600.000 Tweets aus dem Online-Netzwerk Twitter, sowie Daten zu Nutzernamen und einer automatischen Einschätzung der Toxizität der Tweet-Texte zur Verfügung. Es werden zuerst drei Modelle entwickelt, wie sich Interaktionen in dem Netzwerk beschreiben lassen. Anschließend wurden verschiedene Algorithmen angewendet, mit welchen sich Meinungsführer und Gruppen in Graphen bestimmen lassen können. Zur Kontrolle der Ergebnisse wurde ein kurzer Überblick über Meinungsführerschaft gegen und damit die Ergebnisse auf Plausiblität überprüft. Zuletzt wurden mögliche Zusammenhänge zwischen Einfluss-Werten und Toxizitäts-Bewertungen untersucht.

1.1 Verwandte Arbeiten

Zunächst werden einige Arbeiten vorgestellt, welche sich mit Textanalyse in Sozialen Netzwerken, der Analyse von Beziehungen zwischen Nutzern sowie Graphen allgemein beschäftigten. Dies stellt auch die Grundlage dieser Arbeit dar.

Netzwerkanalyse und Social Media Forensik. Ein umfassendes Werk stammt von Al-khateeb u. a. (2019). Sie beschäftigten sich mit Devianz in Sozialen Medien und forensischen Methoden, Beziehungen in Netzwerken aufzudecken und zu analysieren.

Mrsic u. a. (2019) befassten sich mit der Ausbreitung von Informationen in Sozialen Netzwerken, Informationskrieg und der visuellen Darstellung von Netzwerken. Bhat u. a. (2017) entwickelten ein Framework zur Erkennung und Beschreibung von Gruppen in Sozialen Netzwerken. Zweig (2016) beschreibt Algorithmen zur Graphenanalyse und den Ablauf eines solchen Projektes.

Graphentheorie und Graphendatenbanken. Hodler u. a. (2019) und Hodler u. a. (2021) beschrieben die Analyse von Graphen sowie fortgeschrittene Methoden der Graph Data Science. Beide Werke nutzten die Graphdatenbank *Neo4j* als praktisches Anwendungsbeispiel. Merkl Sasaki u. a. (2018) und Robinson u. a. (2015) erklären ebenfalls die Graphendatenbank Neo4j. Zhang u. a. (2020) bot einen Überblick über die Anwendung von Machine Learning auf Graphen.

Meinungsführer. Zuletzt sollen einige Werke über Meinungsführer vorgestellt werden. Diese befassen sich nicht mit Netzwerktheorie, sondern mehr mit der praktischen Anwendung, meist im Kontext von Marketing. Diese Themen werden in dieser Arbeit nur sehr oberflächlich betrachtet. Dressler u. a. (2009) liefert einen umfassenden Überblick über die Forschung über Meinungsführer in den letzten 70 Jahren. Potthoff (2016) stellt eine Zusammenfassung der einflussreichsten Werke zur Meinungsforschung zusammen. Kaiser (2012) und Schach u. a. (2018) beschreiben, wie Unternehmen Influencer für Marketing verwenden können und wie Influencer ihre Position bewerten und ausbauen können.

1.2 Kapitelüberblick

In Kapitel 2 wird genauer auf die benötigten theoretischen Grundlagen zu Forensik, Webscraping und Graphen eingegangen. In Kapitel 3 wird die Anwendung der Grundlagen auf das konkrete Problem beschrieben. In Kapitel 4 werden die möglichen Werkzeuge evaluiert, mit welchen ein solcher Workflow umgesetzt werden könnte und welche Werkzeuge am Ende ausgewählt wurden. In Kapitel 5 werden der Inhalt und die Struktur des untersuchten Datensatzes beschrieben. Kapitel 6 enthält die Umsetzung der Theoretischen Ausarbeitung in das Modell, welches analysiert wird. Kapitel 7 enthält die Auswertung des Datensatzes. In Kapitel 8 werden die Ergebnisse der Arbeit ausgewertet und interpretiert sowie ein Ausblick auf weitere Forschungsthemen gegeben. Anhang A enthält einen Ausblick auf die Möglichkeiten der Anwendung von maschinellem Lernen auf Graphen sowie einen kurzen Vergleich mit den Ergebnissen dieser Arbeit. Anhang B enthält einige Beispiele des geschriebenen Programmcodes.

1.3 Verwendete Software

In dieser Arbeit wurde die Graphendatenbank *Neo4j* verwendet. Die statistische Auswertung erfolgte mit den Python-Paketen *Pandas* (Reback u. a. 2021), *SciPy* (Virtanen u. a. 2020) und *scikit-learn* (Pedregosa u. a. 2011). Die Erstellung der Diagramme erfolgte mit *Matplotlib* (Caswell u. a. 2021) und *seaborn* (Waskom 2021). Die Visualisierung der Graphen erfolgte mit *Gephi* (Bastian u. a. 2009).

2 Grundlagen

In diesem Kapitel werden die Grundlagen zu den Themen Soziale Netzwerke, IT-Forensik, Web Scraping und Graphen erläutert, welche im Rest der Arbeit benötigt werden. Es wird die Bedeutung und Herkunft der Forensik erläutert und wo die Probleme bei der Anwendung auf Soziale Netzwerke liegen. Es werden Möglichkeiten und Hindernisse bei der Sammlung von Daten aus dem Internet kurz vorgestellt und zuletzt einige notwendige Grundlagen über Graphen und ihre Anwendung erläutert.

2.1 Soziale Netzwerke

Soziale Netzwerke bezeichnen generell Beziehungen zwischen Personen und Personengruppen. Die Bedeutung des Begriffs hat sich im Laufe der Zeit verändert und wird heute hauptsächlich auf Webseiten angewendet, auf welchen Personen online miteinander kommunizieren können[1]. Diese Arbeit folgt der Definition des Duden[2]:

Definition 2.1 (Soziales Netzwerk) [Ein] Portal im Internet, das Kontakte zwischen Menschen vermittelt und die Pflege von persönlichen Beziehungen über ein entsprechendes Netzwerk ermöglicht.

In dieser Arbeit wird das soziale Netzwerk Twitter betrachtet. Twitter ist ein Microbloggingdienst des Unternehmens Twitter Inc., welcher 2006 online ging. Das Hauptelement von Twitter sind sogenannte *Tweets*. Ein Tweet ist ein kurzer Text von bis zu 280 Zeichen Länge. Nutzer können vielfältig mit Tweets interagieren. Tweets können beantwortet, retweetet oder quote-tweeted werden. Im ersten Fall erscheint die Antwort in der Konversation direkt unter dem Ursprungstweet. In den beiden anderen Fällen wird eine neue Konversation durch den retweetenden Nutzer erzeugt. Ein Beispiel-Tweet ist in Abb. 2.1 abgebildet. In diesem sind die einzelnen Daten des Tweets in der Weboberfläche von Twitter zu sehen[3].

2.1.1 Filterblasen und Echokammern

Zweig u. a. (2017)

[1] „social network". Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/social20network. Zugegriffen am 18. September 2021.

[2] „Social Network". Dudenredaktion [o. D.], https://www.duden.de/node/167713/revision/167749, zugegriffen am 18. September 2021.

[3] https://twitter.com/Karl_Lauterbach/status/1411729542054617094, zugegriffen am 25.10.2021.

Abbildung 2.1: Ein Beispieltweet. Zu sehen sind (v.o.n.u): Der Anzeigename des Autors, der Twitter-Handle des Autors, der Text des Tweets, ein zitierter Tweet, Datum und Uhrzeit der Erstellung des Tweets, die Anzahl Retweets, die Anzahl zitierter Tweets, die Anzahl „Gefällt mir"-Angaben und die Buttons zum kommentieren, retweeten und Mit „gefällt mir" markieren. Der Tweet enthält keine Hashtags, URLs oder Mentions.

Definition 2.2 (Filterblase) [D]as Phänomen, dass wir von Algorithmen hauptsächlich solche Themen wieder vorgeschlagen bekommen, die wir schon mögen.

Definition 2.3 (Echokammer(n)) Freundesgruppen, die hauptsächlich aus Leuten mit ähnlicher Meinung bestehen, in denen also jede Aussage widerhallt.

Du u. a. (2016) zeigten durch zeitliche Analyse eines Twitternetzwerkes, dass Gruppen auf Twitter sich im Verlauf der Zeit zunehmend polarisieren und die Gruppenstruktur stärker wird. Neu hinzukommende Verbindungen treten deutlich häufiger innerhalb von Gruppen auf als zwischen Nutzern verschiedener Gruppen. Umgekehrt verfallen deutlich mehr Verbindungen zwischen Mitgliedern verschiedener Gruppen als zwischen Mitgliedern der selben Gruppe. Es konnte keine abschließende Erklärung für dieses Verhalten gefunden werden. Eine Möglichkeit ist, dass sich die Nutzer selbst stärker verknüpfen. Eine andere Erkärung ist, dass sich das Online-Netzwerk nur dem realen Netzwerk anpasst, welches bereits stark polarisiert ist.

2.1.2 Meinungsführer und das Zwei-Stufen-Modell der Kommunikation

In der Medienwissenschaft werden Kommunikationsteilnehmer in drei Kategorien unterschieden:

- Opinion leader (Meinungsführer)
- Opinion follower (Ratsuchende, Meinungsfolger) und
- Inaktive

Laut Dressler u. a. (2009, S. 27ff) stellen Meinungsführer die Verbindung zwischen Massenmedien und Ratsuchenden dar (Abb. 2.2). Sie definieren Meinungsführer wie folgt:

Definition 2.4 (Meinungsführer) Der Begriff *Meinungsführer* beschreibt Individuen, die in ihrer Funktion als Bezugspersonen häufig um Rat und um ihre Meinung gefragt werden und dadurch Einfluss auf andere Personen haben. Meinungsführerschaft ist kein Persönlichkeitsmerkmal, sondern eine Verhaltensform, die im Kommunikationsprozess entsteht. Es kann daher nicht dichotom unterschieden werden, ob eine Person ein Meinungsführer ist oder nicht. Es handelt sich vielmehr um eine graduelle Ausprägung.

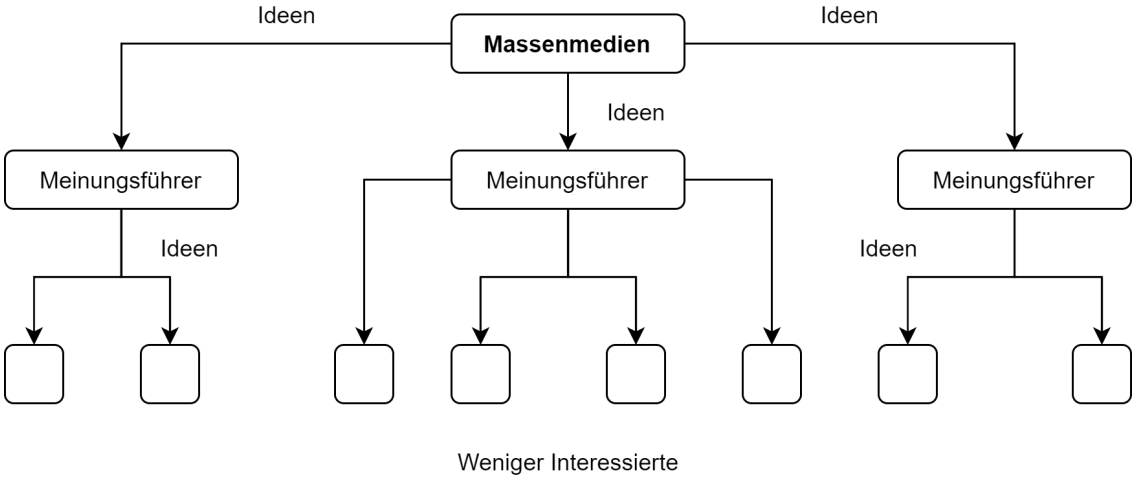


Abbildung 2.2: Die Rolle der Meinungsführer im Zwei-Stufen-Modell der Kommunikation. Sie stehen auf der Zwischenstufe zwischen Massenmedien und Ratsuchenden. Ideen werden durch Meinungsführer gefiltert, interpretiert und anschließend weitergegeben.

Der Fluss der Informationen geht dabei vom Meinungsführer zum Ratsuchenden. Inaktive nehmen kaum oder gar nicht an der Kommunikation teil und werden daher meist nicht betrachtet (Geise 2020).

Die interpersonelle Kommunikation hat einen deutlich stärkeren Einfluss auf die Meinungsbildung als die Massenmedien. Obwohl es Kritik an dem Modell gibt, wurde die

Grundhypothese in wiederholten Studien bewiesen. Einige der Erweiterungen des Modell besagen, dass auch die Kommunikation zwischen Meinungsführern einen bedeutenden Einfluss auf Meinungsführer hat sowie eine Hierarchie zwischen Meinungsführern besteht. Laut Schach u. a. (2018, S. 50) konkurrieren Influencer auf einer Ebene mit Massenmedien um die Konsumenten. Konsumenten lassen sich dabei nach dem Grad ihres Interesses an einem bestimmten Thema in verschiedene Gruppen einteilen.

Meinungsführerschaft entsteht hauptsächlich durch Mitgliedschaften in politischen Vereinigungen oder politisches Engagement. Meinungsführer sind oftmals mit vielen heterogenen Gruppen verbunden (Dressler u. a. 2009, S. 48). Die Netzwerkanalyse wird als aktuellste Methode der Darstellung von Kommunikationsprozessen genannt.

Meinungsführer bestehen immer nur im Kontext ihres sozialen Umfeldes. Sie erkennen einen Informationsbedarf und besitzen die Fähigkeit, diesen Bedarf zu decken. Das Einflusspotenzial des Meinungsführers ist abhängig vom Informationsbedarf des von ihm bedienten sozialen Milieus (Dressler u. a. 2009, S. 60).

Zur Erkennung von Meinungsführern existieren drei grundlegende Möglichkeiten. Diese sind die Selbsteinschätzung, die Befragung von Schlüsselinformanten und die soziometrische beziehungsweise netzwerkanalytische Methode. Letztere ist am genauesten, allerdings am schwierigsten durchzuführen (Dressler u. a. 2009, S. 112).

Durch soziale Netzwerke wird der Aufwand für Netzwerkanalysen extrem verringert. Viele Netzwerke bieten die Möglichkeit, Kommunikationsdaten von Nutzern, Seiten und Diskussionen strukturiert herunterzuladen. Dadurch wird die Auswertung von Daten von tausenden bis Millionen Menschen über sehr lange Zeiträume möglich.

Shafiq u. a. (2013) unterteilen Meinungsführer weiter in *extrovertierte Führer* und *introvertierte Führer*. Extrovertierte Führer interagieren oft mit Freunden, unabhängig von deren Reaktion. Introvertierte Führer interagieren kaum mit Freunden, aber erhalten viele eingehende Interaktionen von Freunden[4].

2.2 IT-Forensik

2.2.1 Überblick

Laut Labudde u. a. (2017, S. 5) beschreibt *Forensik* den Zusammenhang einer Wissenschaft mit dem Rechtssystem. Die digitale Forensik oder IT-Forensik ist ein Teilgebiet der Forensik, das sich besonders mit Computersystemen und digitalen Daten beschäftigt.

[4] Die Einteilung der Gruppen erfolgte durch ein unüberwachtes k-means-Clustering. Dadurch wurden zwei Gruppen gefunden, welche der Beschreibung von Inaktiven und Ratsuchenden entsprechen, auch wenn dieses Modell nicht explizit erwähnt wurde.

Das Bundesamt für Sicherheit in der Informationstechnik (BSI) (BSI 2011, S. 8; BSI 2021, S. 289) definiert IT-Forensik wie folgt:

Definition 2.5 (IT-Forensik) IT-Forensik ist die streng methodisch vorgenommene Datenanalyse auf Datenträgern und in Computernetzen zur Aufklärung von Vorfällen unter Einbeziehung der Möglichkeiten der strategischen Vorbereitung insbesondere aus der Sicht des Anlagenbetreibers eines IT-Systems.

Durch die zunehmende Digitalisierung hat die digitale Forensik bei den Strafverfolgungsbehörden an Bedeutung gewonnen. Während ursprünglich nur Endgeräte wie PCs, Festplatten oder Mobiltelefone für eine Auswertung relevant waren, sind inzwischen auch reine Online-Quellen zunehmend wichtig. Damit sind Webseiten oder Services gemeint, welche nur über das Internet zugänglich sind und weder der Benutzer noch ein forensischer Analyst Zugriff auf die physischen Geräte haben.

2.2.2 Social Media Forensik

Social Media Forensik[5] ist eine Spezialisierung der IT-Forensik mit dem Ziel, Informationen aus Sozialen Netzwerken zu gewinnen. Al-khateeb u. a. (2019, S. 68) definieren den Begriff wie folgt:

Definition 2.6 (Social Media Forensik) Ein Teilgebiet der digitalen Forensik, welcher sich mit Beziehungen zwischen Einheiten (Individuen, Gruppen, Organisationen, etc.) in Sozialen Netzwerken beschäftigt. Verdeckte Beziehungen werden durch Extraktion und Analyse von Metadaten von Social Media Accounts aufgedeckt.

Der Fokus liegt also nicht nur auf der unmittelbaren Sammlung von Daten, wie zum Beispiel Postings in Foren, sondern auch auf der Aufdeckung von verdeckten Beziehungen zwischen mehreren Benutzern. Die verwendeten Metadaten können zum Beispiel Retweets bei Twitter, Likes, Kommentare, IP-Adressen oder Geoinformationen sein.

2.2.3 Der forensische Prozess

Da Forensik immer im Kontext einer Ermittlung geschieht, werden an das Vorgehen des Bearbeiters einige Anforderungen gestellt. So muss das Vorgehen so dokumentiert werden, dass der komplette Vorgang von einem anderen Gutachter oder Richter einwandfrei nachvollzogen werden kann. Zudem sollten standardisierte Prozesse angewendet werden. Im Falle einer Abweichung muss diese begründet werden.

Das einfachste Modell für die Forensik ist das *SAP-Modell* (Abb. 2.3). Das SAP-Modell

[5] In der Literatur je nach Kontext auch *Link Mining*, *Soziale Netzwerkanalyse* oder *Social Cyber Forensics* genannt.

umfasst drei Phasen, *Sicherung*, *Analyse* und *Präsentation*. In der Sicherungsphase werden relevante Beweismittel identifiziert, forensisch gesichert und so vorbereitet, dass sie analysiert werden können. In der Analysephase werden Beweismittel ausgewertet und Schlussfolgerungen formuliert. In der Präsentationsphase werden die gewonnenen Erkenntnisse aufbereitet und dem Gericht vorgestellt. Es wurden auch weitere Modelle

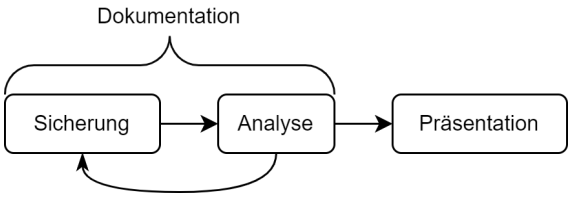


Abbildung 2.3: Darstellung des SAP-Modells mit zusätzlicher Dokumentation der Schritte. (Quelle: eigene Darstellung)

entwickelt, welche das SAP-Modell durch vorbereitende Schritte vor der Sicherung erweitern (BSI 2011) oder die Phasen in weitere Unterphasen unterteilen (Kent u. a. 2006). Diese Arbeit betrachtet nur den Analyse- und Präsentationsschritt des SAP-Modells. Der Sicherungsschritt war nicht notwendig, da die Daten zur Verfügung gestellt wurden.

2.3 Graphentheorie

In diesem Abschnitt werden mathematische Grundlagen von Graphen und ihre Bedeutung für die Social Media Forensik beschrieben.

2.3.1 Überblick

Um ein Netzwerk analysieren zu können, muss es als erstes in ein Format überführt werden, in welchem die gewünschten Funktionen mathematisch modelliert werden können. Für diesen Zweck bietet sich ein Graph an. Graphen wurden bereits beispielsweise von Bhat u. a. (2017), Guarino u. a. (2020), Al-khateeb u. a. (2019) und Mrsic u. a. (2019) genutzt. Ein Graph ist ein mathematisches Modell, in welchem Elemente (genannt *Ecken*) und Beziehungen zwischen diesen Ecken (genannt *Kanten*) dargestellt werden können. Diese Ecken und Kanten können wiederum verschiedene Eigenschaften haben.

Definition 2.7 (Graph) Ein Graph G ist Tupel (V, E), wo V eine Menge der Ecken und E die Menge der Kanten darstellt, welche Ecken miteinander verbinden.

Jede Kante verbindet jeweils zwei Ecken und stellt eine definierte Beziehung dieser beiden Ecken zueinander dar. Wenn Nutzer oder Accounts eines Sozialen Netzwerkes als Ecken betrachtet werden, könnten Kanten zwischen diesen Nutzern zum Beispiel „*ist befreundet mit*" (bei Facebook), „*folgt der Person*" (bei Instagram/Twitter) oder „*hat*

einen Beitrag von Person x geteilt" (Twitter, Facebook) bedeuten. Graphen werden je nach Kontext auch als Netzwerke oder Soziogramme bezeichnet.

2.3.2 Einteilung von Graphen

Graphen können nach mehreren Kriterien eingeordnet werden (Hodler u. a. 2019, S. 18f).

Gerichtet vs. ungerichet: Eine gerichtete Kante hat einen Startpunkt und einen Endpunkt, eine ungerichtete Kante wirkt in beide Richtungen.

Gewichtet vs. ungewichtet: In einem gewichteten Graphen wird die Stärke oder Bedeutung einer Verbindung modelliert.

Monopartiter, bipartiter, n-partiter Graph: Wie viele Partitionen hat der Graph?[6]

Abb. 2.4 zeigen einige der genannten Elemente grafisch.

2.3.3 Labeled Property Graphs

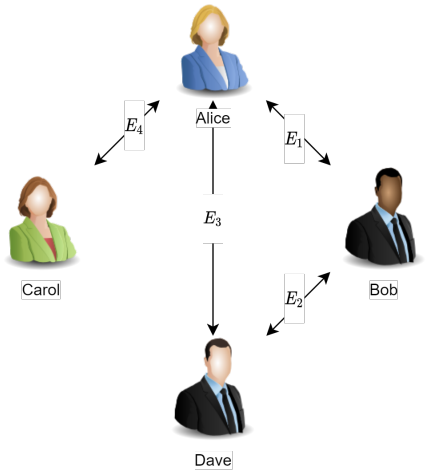
Ein Labeled Property Graph[7] (Webber u. a. 2020, S. 26f) ist eine Variante von Graphen, welche besonders in Graphendatenbanken anzutreffen ist. Sie unterscheidet sich von den rein mathematischen Modellen dadurch, dass der Graph zusätzliche Eigenschaften (Labels, Properties) aufweist. Er besitzt folgende Eigenschaften:

- Er enthält Knoten und Kanten.
- Knoten enthalten Eigenschaften (Schlüssel-Wert-Paare).
- Knoten können ein oder mehrere Labels haben. Labels bezeichnen Gruppen oder Rollen eines Knoten.
- Kanten sind benannt und gerichtet. Sie besitzen immer genau ein Label. Start- und End-Knoten einer Kante müssen im Graph existieren.
- Kanten können ebenfalls Eigenschaften beinhalten.
- Eigenschaften von Knoten und Kanten können als zusätzliche Information für Graphenalgorithmen verwendet werden.

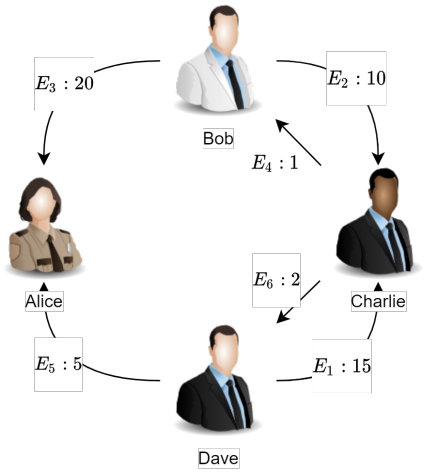
Labeled Property Graphs erlauben es, leicht sehr komplexe Datenmodelle darzustellen. Es können sehr viele unterschiedliche Zusammenhänge in der selben Datenbank gespeichert werden. Mit Hilfe einer geeigneten Sprache (siehe Abschnitt 4.3.1) kann immer nur auf den benötigten Teil der gesamten Informationen zugegriffen, sowie flexibel viele unterschiedliche Informationen verknüpft und manipuliert werden.

[6] Partitionen sind Teilmengen von Knoten eines Graphen, deren Elemente nicht untereinander verbunden sind.

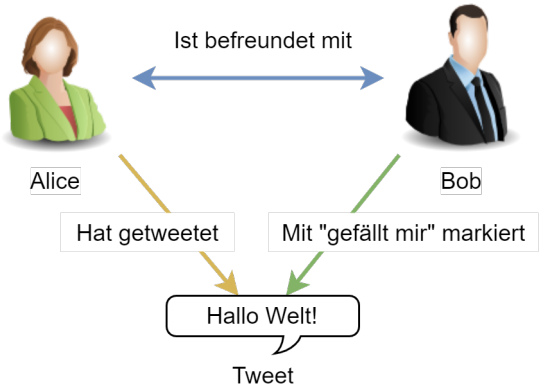
[7] Deutsch: Etikettierter Graph mit Eigenschaften



(a) Einfaches Beispiel eines ungerichteten und ungewichteten Graphen (zum Beispiel Facebook-Freundschaften)



(b) Beispiel eines gerichteten und gewichteten Graph (zum Beispiel „Wie oft hat Person A Person B bei Twitter retweetet.")



(c) Ein heterogener Graph, bei welchem zwei Typen von Ecken und drei Typen von Beziehungen definiert sind.

Abbildung 2.4: Vergleich von gewichteten/ungewichteten und gerichteten/ungerichteten Graphen. (Quelle: eigene Darstellung)

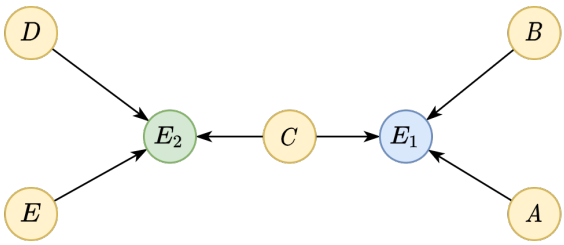


Abbildung 2.5: Beispiel eines Property Graphs. Die blaue und die grüne Ecke stellen die Gruppen und gelbe Ecken die Nutzer dar. Jeder Nutzer ist mit den Gruppen verbunden, in welchen er Mitglied ist.

2.3.4 Graphprojektionen

Viele Graphalgorithmen setzen monopartite Graphen voraus. Real-Life-Datensätzen bestehen aber oft aus bipartiten oder heterogenen Graphen. So kann beispielsweise ein Film-Graph Schauspieler und Filme enthalten. Jeder Schauspieler ist mit den Filmen verbunden, in welchen er mitgespielt hat. Um die Beziehungen von Schauspielern untereinander zu beschreiben, muss eine sogenannte Projektion gefunden werden, Schauspieler direkt zu verbinden. Die entstehenden Verbindungen sind immer ungerichtet und meistens gewichtet. Zweig (2016, S. 137ff) stellt drei Möglichkeiten vor, wie eine monopartite Projektion aus einem bipartiten Graphen erstellt werden kann:

Einfache Projektion: Die simpelste Projektion besteht darin, alle Knoten aus der einen Teilmenge zu verbinden, die mindestens einen gemeinsamen Nachbar aus der zweiten Teilmenge besitzen. Dies würde alle Schauspieler verbinden, welche jemals zusammen in einem Film mitspielten.

Gewichete Projektion mit Schwellwert: Diese Projektion zählt die Anzahl aller Verbindungen zwischen zwei Knoten aus einer Teilmenge. Durch einen Schwellwert können Verbindungen nach Relevanz gefiltert werden. In diesem Fall sagt die Verbindung aus, in wie vielen Filmen zwei Schauspieler zusammen spielten.

Wichtungsschema: Die dritte Möglichkeit ist, ein komplexeres Schema zu verwenden, um das einfache Gewicht zu normalisieren. Dies ist beispielsweise möglich mit dem Jaccard-Koeffizienten, der Okapi BM25-Funktion, der Tangens hyperbolicus-Funktion oder der erwarteten Anzahl von Verbindungen. Diese Projektion enthält die meisten Informationen aus dem ursprünglichen Graphen.

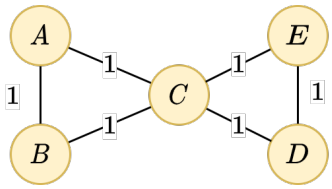


Abbildung 2.6: Abb. 2.5 als Projektion auf einen monopartiten Graphen. Jede der Verbindungen zwischen zwei Nutzern stellt die Anzahl von gemeinsamen Gruppen dar. Eine ähnliche Projektion könnte eine Verbindung der Gruppen durch die Nutzer darstellen.

Es ist theoretisch möglich, bipartite Graphen direkt zu betrachten, allerdings sind diese Methoden schlechter ausgearbeitet als für monopartite Graphen. Wasserman u. a. (1995, S. 326ff) zeigt einige Analysemethoden für bipartite Graphen.

2.3.5 Datenstrukturen für die Speicherung

Es existieren unterschiedliche Möglichkeiten, einen Graphen darzustellen. Diese sind die Inzidenzliste, die Adjatenzliste[8], die Inzidenzmatrix, die Adjatenzmatrix[9] und die Distanzmatrix.

In der Adjatenzmatrix werden die Ecken in einer $n \times n$-Matrix gespeichert, wo n die Anzahl der Ecken ist. Eine 0 bedeutet, dass die Ecken nicht verbunden sind und eine 1, dass sie verbunden sind (Tabelle 2.1). Diese Darstellung wird später genutzt, um den Einfluss der Knoten zu bestimmen.

	Nutzer			
	A	B	C	D
A	0	1	1	1
B	1	0	0	1
C	1	0	0	1
D	1	1	0	0

Tabelle 2.1: Adjatenzmatrix von Abb. 2.4a. (Quelle: eigene Darstellung)

In der Inzidenzliste werden alle Kanten gelistet, zusammen mit ihren dazugehörigen Eckpunkten. In der Adjatenzliste werden alle Eckpunkte mit ihren Kanten gelistet. In der Inzidenzmatrix werden Kanten als Spalten und Ecken als Reihen einer Matrix gespeichert. Eine 0 bedeutet, dass die Ecke nicht auf der Kante liegt und eine 1 bedeutet, dass die Ecke auf der Kante liegt[10]. In Tabelle 2.2 wird der Graph aus Abb. 2.4a in den verschiedenen Listenformen dargestellt. Graphen werden oft in Listenform gespeichert, da diese Form weniger Speicher benötigt. Der vorliegende Datensatz hatte die Form einer Adjatenzliste.

Kante	**Nutzer**
E_1	A, B
E_2	B, D
E_3	A, D
E_4	A, C

(a) Inzidenzliste

Nutzer	**Verbundene Nutzer**
A	B, C, D
B	A, D
C	A
D	A, B

(b) Adjatenzliste

Tabelle 2.2: Listendarstellungen von Abb. 2.4a. (Quelle: eigene Darstellung)

[8] Auch *Nachbarschaftsliste* genannt
[9] Auch *Soziomatrix* oder *Nachbarschaftsmatrix* genannt
[10] Mit -1 kann bei einem gerichteten Graphen angegeben werden, ob die Kante ausgehend oder eingehend ist.

2.3.6 Dateiformate

Graphen können in unterschiedlichen Formaten gespeichert werden. Kanten lassen sich in der Listendarstellung leicht durch einfache Textdateien darstellen, in welcher jede Zeile einen Eintrag einhält. Die Matrixdarstellungen können im `csv`-Format gespeichert werden. Für Ecken eignen sich Formate für strukturierte Daten, wie `json`, `xml` oder ebenfalls `csv`.

2.4 Zusammenfassung

Es wurden die benötigten Konzepte vorgestellt, welche in den folgenden Kapiteln in die Praxis umgesetzt werden. Meinungsführer stellen Informationen für Meinungsempfänger bereit. Diese bezieht sich immer auf einen bestimmten Themenbereich. Soziale Netzwerke sind Plattformen, auf welchen Menschen sich verbinden können. Im forensischen Kontext ist es wichtig, den Ablauf der Sicherung und Analyse genau zu dokumentieren. Daher müssen alle verwendeten Algorithmen erklärt und begründet werden. Diese Verbindungen lassen sich sehr gut durch Graphen darstellen. Dabei müssen komplexe Beziehungen schrittweise vereinfacht werden, bis am Ende eine Darstellung erreicht wird, welche mathematisch analysiert werden kann. Dazu werden Graphenprojektionen genutzt.

3 Analysemethoden der Graph Data Science

In diesem Kapitel werden die fortgeschrittene Konzepte beschrieben, welche auf den Methoden aus Kapitel 2 aufbauen. Nach Hodler u. a. (2019, S. 27 f) und Hodler u. a. (2021, S. 6ff) lassen sich Fragen in der Graphenanalyse in drei beziehungsweise vier Themenbereiche unterteilen:

Bewegung: Dieses Gebiet umfasst die Anwendung von Pfadfindungsalgorithmen. Dies dient dazu, herauszufinden, welche Verbindungen im Netzwerk am relevantesten sind, wie die verschiedenen Netzwerkbereiche miteinander verbunden sind und wo unbekannte Verbindungen liegen. Dieser Schritt ist die Grundlage für die nächsten Themenbereiche.

Einfluss: Beim Einfluss geht es darum, relevante Knoten im Netzwerk zu finden. *Relevanz* kann dabei unterschiedlich definiert sein. Sie kann zum Beispiel beschreiben, welche Knoten Brücken zwischen Teilnetzen darstellen, Senken oder Ausgang von vielen Verbindungen oder Flaschenhälse sind. Relevanz wird in der GDS als *Zentralität* bezeichnet. Diese Algorithmengruppe wird dazu benutzt, Meinungsführer zu bestimmen.

Gruppen: Gruppen sind essenziell um Dynamiken in einem Netzwerk zu verstehen. Allgemein wird eine Gruppe dadurch definiert, dass ihre Mitglieder untereinander mehr Verbindungen haben als die Gruppe zum Rest des Graphen. Nach der Bestimmung einzelner Gruppen können die Miglieder auf Gemeinsamkeiten und verschiedene Gruppen auf Unterschiede untersucht werden. Gruppenzugehörigkeiten werden auch häufig zur Visualisierung genutzt.

Mustererkennung: Die Mustererkennung versucht, in dem Graphen komplexere Zusammenhänge als reine Statistiken zu finden. Diese Methoden bauen auf der Gruppenerkennung und Zentralitätsbestimmung auf. An dieser Stelle wird teilweise mit Methoden des maschinellen Lernens gearbeitet.

In dieser Arbeit wird der Schwerpunkt auf der Ausarbeitung und dem Vergleich von Einfluss in einem Twitter-Netzwerk liegen.

3.1 Einfluss von Knoten

Aus dem Netzwerk sollen die Nutzer bestimmt werden, welche den größten Einfluss auf das Netzwerk haben. Dazu muss zuerst *Einfluss* definiert werden und anschließend Methoden entwickelt werden, wie dieser Einfluss bestimmt werden kann. Der Duden[11] definiert Einfluss als

[11] „Einfluss". Dudenredaktion [o. D.], https://www.duden.de/node/37590/revision/37619, zugegriffen am 18. September 2021.

- beeinflussende, bestimmende Wirkung auf jemanden, etwas; Einwirkung
- Ansehen, Geltung
- Synonyme: Beeinflussung, [Ein]wirkung, Achtung, Ansehen

In dieser Arbeit wird mit folgender Definition von *Einfluss* gearbeitet:

Definition 3.1 (Einfluss) Der Einfluss eines Nutzers auf das Netzwerk ist die Fähigkeit des Nutzers, Reaktionen von anderen Nutzern hervorzurufen. Je weniger eigene Aktivität ein Nutzer aufweist im Verhältnis zu den erhaltenen Reaktionen, desto größer ist der Einfluss jeder einzelnen Aktivität.

Diese Definition soll wiederspiegeln, dass *Meinungsführer* in einem Online-Netzwerk nicht immer nur kompetent oder bekannt sind. Meinungsführer können auch mit sehr kontroversen oder sogar negativen Beiträgen die Stimmung einer Online-Diskussion beeinflussen.

In der Graphentheorie werden *Zentralitätsmaße* verwendet, um die Wichtigkeit eines Knotens zu bewerten. Zweig (2016, S. 245) definiert Zentralität wie folgt:

Definition 3.2 (Zentralität) Die geringste Gemeinsamkeit aller Zentralitätsindizes ist, dass ein Zentralitätsindex eine reellwertige Funktion auf den Knoten eines Graphen ist, d.h. er weist allen Knoten eine reelle Zahl zu. Dieser Wert ist nur abhängig von der *Struktur* des Graphen, nicht von externen Parametern, die den Knoten zugeordnet sind. Je höher der Zentralitätsindex des Knoten, desto höher ist die wahrgenommene Zentralität des Knoten im Netzwerk. Damit wird durch die Zentralität eine Ordnung der Knoten definiert. [Hervorhebung im Original]

Diese Definition ist sehr schwach, aber ausreichend genau im Zusammenhang mit der Definition des Duden sowie den Bedingungen für Einfluss in sozialen Netzwerken. Durch den numerischen Wert der Zentralität kann der Grad der Meinungsführerschaft gemessen werden. Dafür muss zuerst eine geeignete Methode gefunden werden, die Nutzer zu strukturieren.

In der Praxis werden Daten der Knoten oftmals benutzt, um die Stärke der Beziehung zu anderen Knoten zu berechnen. Dadurch werden sie in die Struktur des Netzwerkes einbezogen.

Es existieren sehr viele verschiedene Methoden, die Zentralität zu bestimmen. Jede Methode bewertet einen bestimmten Aspekt des Graphens und setzt implizit gewisse Annahmen über die Funktionsweise des Netzwerkes vorraus. Daher muss ein Zentralitätsalgorithmus gewählt (oder definiert) werden, welcher das Netzwerkmodell am Besten abbildet (Zweig 2016, S. 445). Umgekehrt erlaubt die Anwendung von verschiedenen Zentralitätsalgorithmen, begründete Hypthesen über die (unbekannte) Funktionsweise eine Netzwerkes zu formulieren (Zweig 2016, S. 457). Saxena u. a. (2020), Wan u. a.

(2020) und Xu u. a. (2020) geben einen umfassenden Überblick über verschiedene Zentralitätsmaße.

Grad. Die grundlegendste Möglichkeit der Berechnung ist der *Grad*. Der Grad $deg(v)$ einer Ecke v ist die Anzahl an Ecken, welche mit dieser Ecke verbunden sind. Bei gerichteten Graphen lässt sich dies noch in *Eingangsgrad* $deg_{in}(v)$ und *Ausgangsgrad* $deg_{out}(v)$ unterscheiden, für die Anzahl eingehender und ausgehender Kanten.

Die Grad-Zentralität erlaubt keine Aussage über die Relevanz der einzelnen Verbindungen. Es wird davon ausgegangen, dass ein Knoten mit vielen Verbindungen sehr wichtig ist.

PageRank. Der PageRank (Brin u. a. 1998) wurde für die Suchmaschine Google entwickelt. Er beschreibt ein Modell, bei welchem ein Internet-Nutzer auf einer Seite startet und zufällig auf eine von dieser Seite verlinkte Seite wechselt. Dabei wird der PageRank der Seite gleichmäßig an alle verlinkten Seiten weitergegeben. Eine Seite ist damit einflussreich, wenn viele andere einflussreiche Seiten auf sie verweisen. Die Formel lautet

$$C_P(v) = \frac{\alpha}{N} + (1-\alpha) \sum_{t \in V} \mathbf{A}_{v,t} \frac{C_P(t)}{deg_{out}(t)}$$

wo

- N die Gesamtzahl aller Ecken im Netzwerk,
- $\mathbf{A}$ die Adjatenzmatrix und
- α eine Konstante im Intervall $[0, 1]$

bezeichnen. α ist die Wahrscheinlichkeit, dass der Nutzer tatsächlich auf eine verlinkte Seite wechselt. Mit einer Wahrscheinlichkeit von $1 - \alpha$ wechselt der Nutzer zu einer zufälligen Seite. Es wird oft $\alpha = 0.85$ als Standard gewählt.

3.2 Einfluss in der Sozialen Netzwerkanalyse

In den Naturwissenschaften sind die Eigenschaften einzelner Knoten oft irrelevant, daher werden ausschließlich statistische Werte des Netzwerkes betrachtet, aber keine einzelnen Knoten. In den Sozialwissenschaften dagegen werden oftmals einzelne Knoten ausgewertet (Zweig 2016, S. 39f). In der SNA wird besonderer Wert darauf gelegt, Meinungsführer zu finden, welche einen besonders großen Einfluss ausüben. *Besonders großer Einfluss* wird durch drei Kriterien ausgezeichnet (Katz 1957, S. 72):

1. Personifizierung bestimmter Werte (Wer bin ich?)

2. Kompetenz (Was kann ich?)
3. Strategische Soziale Position (Wen kenne ich?)

Dafür wurden weitere Algorithmen entwickelt, welche besondere Eigenschaften von Beziehungen in sozialen Netzwerken abbilden sollen. Dazu zählen der LeaderRank (L. Lu u. a. 2011), TwitterRank (Weng u. a. 2010), ArticleRank (Li u. a. 2009) und @Rank (Lubarski u. a. 2014). Die bisher genannten Algorithmen wie PageRank und LeaderRank zeigen Schwächen bei dieser Aufgabe, insbesondere wenn das Netzwerk eine stern-ähnliche Topologie besitzt. Dafür wurde der CompetenceRank (Spranger u. a. 2020) entwickelt. Lyu u. a. (2021) zeigten, dass Gruppenzugehörigkeiten genutzt werden können, um Einfluss zu bestimmen. Rajeh u. a. (2021) zeigten, dass Zentralitätsmaße, welche Gruppen beachten, deutlich von „klassischen" Zentralitätsmaßen abweichen können, besonders, wenn eine starke Gruppenstruktur vorliegt.

LeaderRank: Erweiterung des PageRanks mit einem Grund-Knoten für bessere Ranking-Effizienz und Robustheit gegen Rauschen.

ArticleRank: Eine Variante des PageRanks, welche Knoten mit hohem Ausgangsgrad stärker bestraft.

TwitterRank: Eine Erweiterung des PageRanks, welche thematische Ähnlichkeiten zwischen Nutzern betrachtet. Tweets eines Nutzers werden unüberwacht einem Thema zugeordnet. Die Wahrscheinlichkeit, dass ein Nutzer einem anderen Nutzer folgt, ist proportional zur Ähnlichkeit der Tweet-Themen der Nutzer.

@Rank: Für E-Mail-Verkehr wird die Zeitdifferenz zwischen dem Erhalt der Mail und dem Versand der Antwort bestimmt. Je kürzer diese Zeit ist, desto wichtiger ist der Empfänger.

DiverseCentrality: Eine Variante des PageRanks, bei welcher Knoten, welche mehreren Gruppen angehören, relevanter sind als Knoten, welche nur einer Gruppe angehören.

CompetenceRank: Eine Erweiterung des LeaderRanks, welche besonders bei Netzwerken mit Stern-ähnlicher Topologie bessere Ergebnisse erzielt und Informationen wie Likes und Retweets mit einbezieht.

In dieser Arbeit wurden die Ansätze von Li u. a. (2009) und Spranger u. a. (2020) verknüpft, um die Beziehungen zwischen Nutzern durch die Eigenschaften der einzelnen Tweets zu bewerten.

ArticleRank. Der ArticleRank ist eine Variante des PageRanks, welche der Annahme folgt, dass Verbindungen von Knoten mit geringem Grad eine höhere Bedeutung besitzen als Verbindungen von Knoten mit hohem Grad. Dafür wird der Einfluss von Knoten mit hohem Grad in jeder Iteration verringert. Er wurde ursprünglich für Zitatnetzwerke von wissenschaftlichen Journalen entwickelt. Durch seine Eigenschaft, Knoten mit hohem

Ausgangsgrad zu bestrafen, eignet er sich auch zur Bewertung von Personen in Sozialen Netzwerken nach Definition 3.1. Er berechnet sich nach:

$$C_{AR}(v) = (1-d) + d * \overline{deg_{out}} * \sum_{w \in N_{in}(v)} \frac{C_{AR}(w)}{deg_{out}(w) + \overline{deg_{out}}}$$

wo

- $\overline{deg_{out}}$ den durchschnittlichen Ausgangsgrad des Graphen,
- $N_{in}(v)$ eingehende Nachbarn von v und
- d einen Dämpfungsfaktor im Intervall $[0,1]$

bezeichnet. Ähnlich wie beim PageRank wird meist mit $d = 0.85$ als Standardwert gearbeitet.

Einfluss von Knoteninformationen In Sozialen Netzwerken stehen oftmals nicht nur die rein topologischen Informationen zu Verfügung, sondern weitere Daten, anhand derer ein Post bewertet werden kann. So können Beiträge mit „Gefällt mir" markiert, geteilt oder kommentiert werden. Diese Informationen sind direkte Indikatoren dafür, wie einflussreich ein Beitrag ist. Ein Zentralitätmaß für Soziale Netzwerke sollte daher diese Informationen mit einbeziehen. Der CompetenceRank ist der einzige Algorithmus, welcher diese Informationen betrachtet. Dazu werden die Anzahl „Gefällt mir"-Angaben und *Geteilte Posts*[12] mit der Postfrequenz eines Nutzers verrechnet. Dieser Ansatz wird in Kapitel 6 aufgegriffen.

3.3 Erkennung von Gruppen

Neben der Frage, wie die Wichtigkeit eines Knoten im Netzwerk bestimmt werden kann, ist relevant, welche Ecken des Graphen ähnlich zueinander sind. Informell wird eine Gruppe (von Menschen) wie folgt definiert[13].

- kleinere Anzahl von [zufällig] zusammengekommenen, dicht beieinanderstehenden oder nebeneinandergehenden Personen [die als eine geordnete Einheit erscheinen]
- Gemeinschaft, Kreis von Menschen, die aufgrund bestimmter Gemeinsamkeiten zusammengehören, sich aufgrund gemeinsamer Interessen, Ziele zusammengeschlossen haben.

[12] Das Facebook-Gegenstück zu einem Retweet bei Twitter.

[13] „Gruppe", Dudenredaktion [o. D.], https://www.duden.de/node/60970/revision/61006, zugegriffen am 18. September 2021.

Gruppen können also sowohl bewusst als auch zufällig zusammenkommen.

In der Graphentheorie wird diese Definition genauer spezifiziert. Mitglieder einer Gruppe sind stärker untereinander vernetzt als mit Nutzern aus anderen Gruppen. In der SNA können somit Freundeskreise, Arbeitsgruppen oder politische Gruppen gefunden werden.

Definition 3.3 (Gruppe) Eine Gruppe ist eine Teilmenge von Ecken in einem Graphen, welche untereinander stärker verknüpft sind als mit anderen Ecken aus dem selben Graphen (vgl. Shin u. a. 2014, S. 568; Hodler u. a. 2019, S. 115).

Gruppenerkennung wird des Weiteren genutzt, um Mitglieder einer Gruppe zu einzelnen Knoten zusammenzufassen und damit intra- und inter-Gruppen-Beziehungen analysieren zu können. Es gibt mehrere Möglichkeiten, solche stark verbundenen Gruppen zu finden.

Modularität. Um diese stark zusammenhängenden Ecken zu finden, wird die Modularität verwendet (H. Lu u. a. 2014). Die Modularität Q eines Graphen ist definiert als

$$Q = \frac{1}{4m} \sum_{i,j} \left(\mathbf{A}_{i,j} - \frac{k_i k_j}{2m} \right) * \delta(c_i, c_j)$$

wo

- i, j Knoten,
- $\mathbf{A}$ die Adjatenzmatrix,
- k_i, k_j die Summe der Gewichte der an i und j angebundenen Kanten,
- m die Summe aller Kantengewichte des Graphen,
- c_i, c_j die Gruppe von i und j und
- $\delta(c_i, c_j)$ das Kronecker-Delta ($\delta(c_i, c_j) = 1$ wenn $c_i = c_j$, sonst 0.)

bezeichnen.

Mit dieser Messzahl kann bestimmt werden, wie stark ein Graph von der erwarteten zufälligen Verteilung von Kanten abweicht. Bei der modularitätsbasierten Clustererkennung wird versucht, Teilgraphen von G zu finden, welche intern eine größere Modularität als G besitzen, aber zwischen den Clustern eine geringere Modularität zu finden ist.

3.4 Ähnlichkeitsbestimmung

Die Ähnlichkeitsbestimmung hat das Ziel, paarweise Beziehungen zwischen allen Knoten zu berechnen. Dies wird oft als Vorverarbeitung genutzt, um Projektionen von bipartiten Graphen auf monopartite Graphen zu erzeugen. Ähnlichkeit (bzw. Similarität) bedeutet:

- „Ähnlichkeit mit etwas anderem, sichtbar daran, dass diese zwei in mindestens einem Aspekt Gleichheit und mindestens einem weiteren Ungleichheit aufweisen“[14]
- „eingeschränkte Gleichheit/ Übereinstimmung“[15]

Ähnlichkeit von zwei Knoten in einem Graphen kann bedeuten, dass diese Knoten selbst verbunden sind. Es kann aber auch bedeuten, dass diese Knoten viele gemeinsame Nachbarn besitzen. Die strukturelle Ähnlichkeit ist der strengste Ähnlichkeitsbegriff. Zweig (2016, S. 187) definiert sie wie folgt:

Definition 3.4 (Strukturelle Ähnlichkeit) Eine strukturelle Knotenähnlichkeit weiß nichts über die durch einen Knoten repräsentierte Entität; sie quantifiziert die Ähnlichkeit zwischen den Verbindungsmustern zweier Knoten. Die Hauptannahme hinter der Messung der strukturellen Ähnlichkeit besteht darin, dass eine Ähnlichkeit der beiden durch die Knoten repräsentierten Entitäten abgeleitet werden kann.

Jaccard-Ähnlichkeit. Die Jaccard-Ähnlichkeit bestimmt die Überschneidung zweier Mengen. Das geschieht, indem für alle Knoten paarweise die benachbarten Knoten betrachtet werden. Dabei wird das Verhältnis aus der Größe der Schnittmenge und der Größe der Vereinigungsmenge der Nachbarmengen gebildet. Dies eignet sich besonders, um bei bipartiten Graphen Ähnlichkeiten zwischen Knoten einer Partition zu bestimmen. Die Formel lautet:

$$J(v_i, v_j) = \frac{|N_{out}(v_i) \cap N_{out}(v_j)|}{|N_{out}(v_i) \cup N_{out}(v_j)|} = \frac{|N_{out}(v_i) \cap N_{out}(v_j)|}{|N_{out}(v_i)| + |N_{out}(v_j)| - |N_{out}(v_i) \cap N_{out}(v_j)|}$$

mit $v_i \neq v_j$, wo $N_{out}(v)$ ausgehende Nachbarn von v bezeichnen.

3.5 Weitere Algorithmen

Neben den bisher genannten Methoden existieren auch noch weitere Möglichkeiten, Graphen zu analysieren. Dazu zählen Node Embeddings, Link Prediction und Maschine Learning.

Node/Graph Embeddings: Note Embeddings haben das Ziel, Knoten und ihre Eigenschaften auf Vektoren mit geringer Dimension abzubilden, so dass die relativen Abstände der Knoten zueinander erhalten bleiben. Dies wird oft als Vorverarbeitung für den Einsatz von Machine Learning genutzt. Graph Embeddings sollen es ermöglichen, verschiedene Graphen miteinander zu vergleichen. *Node2vec* von

[14] „Similarität”, Wikitionary, Zugegriffen am 28. November 2021, https://de.wiktionary.org/wiki/Similarität

[15] „Ähnlichkeit”. Dudenredaktion [o. D.], https://www.duden.de/node/2994/revision/3020. Zugegriffen am 21. September 2021.

Grover u. a. (2016) ist ein aktuelles Beispiel, welches sehr gute Ergebnisse beim Node Embedding erreicht.

Link Prediction: Link Prediction hat das Ziel, zukünftige Verbindungen vorherzusagen oder fehlende, bereits bestehende Verbindungen aufzudecken. Dazu werden strukturelle Elemente wie die Dreiecksanzahl bewertet.

Machine Learning: Maschine learning ist eine Methode, die für fast alle Graphenfragen angewendet werden kann. Mit Machine Learning können Knoten und Verbindungen klassifiziert, bewertet oder vorhergesagt werden. Es werden Node Embeddings benutzt, um Daten zu erhalten, mit welchen beispielsweise künstliche neuronale Netze oder Support Vector Machines trainiert werden können.

Machine Learning ist zu umfangreich für diese Arbeit und wird daher nicht Teil der Analysen sein. In Anhang A wird einn Beispiel gezeigt, wie node2vec anwendbar ist. Bronstein (2020) und Zhang u. a. (2020) bieten gute Informationen über Maschinelles Lernen mit Graphen.

3.6 Kritik

Zweig (2016, S. 468) argumentiert, dass eine gute theoretische Erklärung einer Netzwerkanalyse nicht automatisch bedeutet, dass diese Analyse richtig ist. Zur Bewertung der Ergebnisse werden zwei Möglichkeiten genannt:

1. A priori-Wissen auf Basis der untersuchten Hypothese und
2. Die post-hoc-Analyse.

Im ersten Fall wird vor der Analyse bestimmt, wie das gewünschte *richtige* Ergebnis auszusehen hat oder das Ergebnis wird mit anderen Modellierungsmethoden verglichen und auf Plausibilität geprüft. Im zweiten Fall wird das Ergebnis auf Muster überprüft, welche nicht a priori bekannt oder spezifiziert waren. Oftmals führt eine post-hoc-Analyse zu einer veränderten Hypothese oder neuen Experiementen. Agarwal u. a. (2019, S. 234) bezeichnen Rohdaten als *Rohöl*, welches erst raffiniert werden muss, um daraus sinnvolle Schlüsse zu ziehen. Für diese Raffination wird umfangreiches Domain-Wissen benötigt.

Meinungsführerschaft ist eine sehr interdisziplinäre Forschung. In der Literatur wird die wird die Unterscheidung zwischen naturwissenschaftlichen und soziologischen Ansätzen deutlich. Zur Frage, mit welchen technischen Methoden Meinungsführer gefunden werden können, existiert eine große Bandbreite an Forschung. Die Analyse dieser Algorithmen bezieht sich dabei allerdings oft rein auf mathematische beziehungsweise technische Aspekte, wie statistische Verteilungen, Berechnungseffizienz und die Robustheit gegen Rauschen.

Bei der Suche nach Meinungsführern ist es möglich, das Ergebnis anhand anderer Kriterien zu verifizieren. Es wäre möglich, zu untersuchen, welche Bedeutung die gefundenen Meinungsführer politisch oder gesellschaftlich haben. Ausgehend von der rein mathematischen Beschreibung eines Meinungsführers könnten Rückschlüsse auf die Kriterien von Katz sein: Welche Werte personifizieren die Meinungsführer? Welche Kompetenzen besitzen sie? Welche soziale Position haben sie inne?

3.7 Zusammenfassung

In diesem Kapitel wurde ein Überblick über Probleme und Fragen in der Graphentheorie gegeben. Dazu wurden Algorithmen vorgestellt, welche für diese Probleme verwendet werden. Besonderer Fokus lag dabei auf der Zentralitätsbestimmung. Zentralität ist das Maß, mit dem in der SNA Meinungsführer bestimmt werden. Es von wurden acht Algorithmen der ArticleRank als am besten geeignet ausgewählt. Für die Bestimmung der Ähnlichkeit von Nutzern wurde die Jaccard-Ähnlichkeit ausgewählt.

4 Recherche und Auswahl der Werkzeuge

Für eine Graphenanalyse wird passende Software benötigt. Dafür wurde ein Vergleich verschiedener Tools durchgeführt, welcher im Folgenden erläutert wird. Zuerst wurde eine Liste von Kriterien definiert, welche überprüft wurden und die Ergebnisse vorgestellt.

4.1 Kriterien

Um Netzwerke in einem Programm bearbeiten zu können ist es notwendig, diese auch in einem einfach zu verarbeitenden Format speichern zu können. Für die Analyse der Graphen ist es wichtig, dass verschiedene Dateiformate importiert und auch zur Weiterverarbeitung exportiert werden können. Standardformate für strukturierte Daten umfassen zum Beispiel .csv oder .json, es existieren aber auch Formate speziell für Graphen. Um Graphen zu manipulieren, müssen Kanten und Ecken einfach hinzugefügt und entfernt werden können. Außerdem müssen Kanten und Ecken Daten speichern können, welche ebenfalls verändert werden können. Die Analysemöglichkeiten von Graphen wurden in Kapitel 3 erläutert. Das Werkzeug sollte eine Vielzahl der dort genannten Algorithmen unterstützen und sich erweitern und konfigurieren lassen.

Neben den mathematischen Analysen des Netzwerkes ist es ebenso wichtig, die Ergebnisse auch darstellen zu können. Dies ist wichtig, um ein intuitives Verständnis für die Daten zu erlangen sowie die Ergebnisse besser interpretieren zu können. Die einfachste Möglichkeit einen Graphen darzustellen besteht darin, Ecken durch Punkte oder Kreise und die Verbindungen durch Linien zwischen diesen Kreisen darzustellen. Sehr viele weitere Eigenschaften des Graphen können durch Parameter der Darstellung visualisiert werden. So können Pfeile die Richtung und die Dicke des Strichs die Gewichtung einer Kante darstellen. Ecken können unterschiedlich groß oder unterschiedlich gefärbt dargestellt werden, womit zum Beispiel Zentralität und Gruppe kodiert werden können. Durch die Anordnung der einzelnen Ecken können weitere zugrinde liegende Zusammenhänge dargestellt werden, indem Gruppen nahe beieinander dargestellt und räumlich von anderen Gruppen abgegrenzt werden.

Kategorie	Beschreibung
Import und Export	In welchen Formaten kann das Werkzeug Graphen einlesen und ausgeben?
Analyse	Wie viele der Analysemethoden aus Kapitel 3 für Pfadfindung, Zentralität und Gruppenerkennung werden unterstützt? Gibt es sonstige interessante Funktionalitäten?
Visualisierung	Bietet das Werkzeug Möglichkeiten, Graphen optisch darzustellen und diese Darstellung anzupassen?
Erweiterbarkeit	Ist es möglich, nicht vorhandene Funktionalität manuell hinzuzufügen? Gibt es die Möglichkeit, Erweiterungen von anderen Entwicklern hinzuzufügen?
Aktueller Zustand	Wie ausgereift ist das Werkzeug? Ist es weit verbreitet und wird aktuell noch entwickelt? Läuft es stabil?
Zukünftiger Support	Wird das Werkzeug auch in Zukunft verwendbar sein? Gibt es Support, welcher bei Problemen reagiert?
Dokumentation	Wie ausführlich ist das Werkzeug dokumentiert? Existieren Tutorials oder Beispiele für neue Programmierer?

Tabelle 4.1: Kriterien, nach welchen die verwendeten Programme ausgesucht wurde.

4.2 Vergleich der Werkzeuge

4.2.1 Native Unterstützung von Programmiersprachen

Am Anfang soll überprüft werden, ob Programmiersprachen existieren, welche Graphen nativ verarbeiten können. Dies ist nicht der Fall. Um eine solche Funktionalität zu schreiben ist eine analytische Programmiersprache wie *Python* oder *R* vorzuziehen. Für diese gibt es viele Pakete oder Bibliotheken, welche derartige Funktionen bieten. Bei Python wird empfohlen, eine Inzidenzliste als Python-Liste von Python-Sets oder Dictionaries zu implementieren (vgl. van Rossum 2019). Bei universellen Programmiersprachen kann zudem theoretisch jede beliebige Funktionalität implementiert werden. Sowohl Python als auch R sind gut etablierte und verbreitete Programmiersprachen. Beide sind also auch langfristig verlässliche Werkzeuge. Beide Programmiersprachen können .csv und .json lesen und verarbeiten, aber weitere Formate werden nicht unterstützt. Aufgrund der mangelnden Unterstützung und der Komplexität der gewünschten Funktionen sollte dies daher nur die letzte Möglichkeit sein, wenn keine andere Software oder Bibliotheken zur Verfügung stehen. Es wurden Python in der Version 3.9.2 und R in der Version 4.0.2 betrachtet.

4.2.2 Programm-Bibliotheken

Es existieren fortgeschrittene Bibliotheken zur Verarbeitung von Graphen. Diese zum Beispiel umfassen *SNAP* (Leskovec u. a. 2016), *NetworkX* (Hagberg u. a. 2008) und *igraph* (The igraph core team 2021). Diese können Graphen auch visualisieren und in einer Reihe von unterschiedlichen Formaten exportieren, um sie anderweitig zu verarbeiten. Die drei genannten Bibliotheken wurden untersucht, da sie für Python zur Verfügung stehen. igraph kann zusätzlich noch mit *C*, *Matlab* und *R* verwendet werden. SNAP ist außerdem verfügbar für *C++*. Alle Bibliotheken ist kostenlos verfügbar.

4.2.3 Graphentools

Eigenständige Programme um mit Graphen zu arbeiten sind zum Beispiel *Gephi* (Bastian u. a. 2009). Gephi ist ein Programm, welches primär auf die Visualisierung von Graphen spezialisiert ist. Es bietet mehrere Importmöglichkeiten sowie die Möglichkeit, Statistiken wie Zentralität oder Partitionierung zu berechnen. Es werden viele mächtige Funktionalitäten und Möglichkeiten der Darstellung geboten. Allerdings stammt die letzte Version von 2017 und es findet keine nennenswerte Entwicklung mehr statt. Daher ist fragwürdig, wie lange es noch verwendbar ist. Aus diesem Grund wird es nicht weiter in dieser Arbeit betrachtet. Gephi ist kostenlos verfügbar. Es wurde in der Version 0.9.2 untersucht.

4.2.4 Datenbanksysteme

Eine weitere Möglichkeit der Umsetzung besteht in einem Datenbanksystem. Dafür gibt es spezielle Systeme für Graphendatenbanken, wie zum Beispiel Neo4j (Neo4j, Inc. 2021). Als spezialisierte Graphen-Datenbank stehen umfangreiche Methoden zur Verfügung, um Graphen zu manipulieren und zu analysieren. Neben den integrierten Funktionen kann mit der Abfragesprache Cypher frei auf den Graphen zugegriffen werden. Die Datenbank ist sehr ausgereift, allerdings befinden sich einige der fortgeschrittenen Analysemethoden und Machine Learning-Algorithmen noch im Alpha- oder Betastadium. Außerdem bestehen Schnittstellen zu vielen üblichen Programmiersprachen. Umfangreiche Dokumentation von Neo4j ist in (Hodler u. a. 2019) und (Hodler u. a. 2021) zu finden. Zur Visualisierung liefert Neo4j das Tool *Bloom*, mit welchem Graphen dargestellt werden können. Zusätzlich können die Tools *GraphXR*, *Graphlytic Desktop* und *NeoMap Map Visualizer* im eigenen App Manager installiert werden. Es wurde Neo4j in der Version 4.2.6 betrachtet. Die Community-Edition ist kostenlos verfügbar, für professionellen Einsatz wird die kostenpflichtige Enterprise-Edition mit erweiterten Funktionen angeboten.

4.2.5 Auswahl des Werkzeuges

Neo4j bietet in allen drei Kriterien umfangreiche und gut dokumentierte Funkionen und scheint durch seinen Umfang und die Verbreitung auch langfristig nutzbar zu sein. Daher wurde entschieden, die Umsetzung mit Neo4j durchzuführen. Aufgrund dieses Vergleichs wurde entschieden, das Datenbanksystem Neo4j zur Umsetzung zu verwenden.

4.3 Neo4j

Für die praktische Umsetzung der Analyse wurde die Graphendatenbank Neo4j ausgewählt. Im Folgenden wird das System kurz vorgestellt.

4.3.1 Cypher

Cypher ist die Programmiersprache für Neo4j. Sie ist semantische Abfragen auf Graphen ausgelegt und wurde an der Sprache SQL für relationale Datenbanken angelehnt. Die Syntax von Cypher erlaubt die Beschreibung von Mustern in einem Stil ähnlich ASCII-Art[16]. In Cypher wird beschrieben, *was* getan werden soll, aber nicht *wie* es getan werden soll. Die genaue Interpretation und Ausführung übernimmt der Compiler (Neo4j, Inc, openCypher 2021).

Knoten werden in Cypher durch runde Klammern gekennzeichnet: `()` oder `(p)`. Durch einen Variablennamen in der Klammer kann später auf diesen Knoten zugegriffen werden. Labels werden durch Doppelpunkte angegeben: `(p:Person)`. Diese stellen Rollen oder Typen dar. Knoten können Eigenschaften haben. Diese werden durch Schlüssel-Wert-Paare angegeben: `(p:Person {Name: "Sandro"})`.

Beziehungen werden durch Bindestriche oder eckige Klammern in Bindestrichen geschrieben: `-->` oder `-[s:STUDIERT]->`. Die Richtung der Beziehung wird durch das „>"- beziehungsweise „<"-Zeichen angegeben. Dies soll einen Pfeil darstellen. Beziehungen müssen bei der Erstellung immer eine Richtung zugewiesen bekommen, können aber bei Abfragen in beide Richtungen betrachtet werden. Beziehungen können ebenfalls Eigenschaften besitzen: `-[:STUDIERT {Beginn: "2015"}]->`. Eine sehr einfache Abfrage ist in Listing 4.1 gezeigt.

4.3.2 Aufbau der Datenbank

Die Grundlage aller Analysen in Neo4j ist die Datenbank. Diese umfasst die Rohdaten auf der untersten Abstraktionsschicht. Allerdings wird im Normalfall nicht die gesamte Daten-

[16] Eine Kunstrichtung, die mit Buchstaben, Ziffern und Sonderzeichen einer nichtproportionalen Schrift kleine Piktogramme oder ganze Bilder darzustellen versucht.

```
1 MATCH (p:Person)-[:STUDIERT_IN {Beginn: "2015"}]->(:Stadt {
    Name: "Mittweida"})
2 RETURN p.name
```

Listing 4.1: Beispielabfrage in Cypher. Diese gibt die Namen aller Studenten der Hochschule Mittweida zurück, welche 2015 anfingen zu studieren.

bank zur Auswertung betrachtet. Stattdessen wird ein *Graph* projiziert. Ein Neo4j-Graph ist eine Teilmenge der gesamten Datenbank. Diese wird für verbesserte Performance komplett im RAM gehalten. Es ist zu beachten, dass bei Neo4j immer zwischen *Graph* und *Datenbank* unterschieden werden muss. Es existieren zwei Arten von Graphen:

Benannte Graphen: Ein bestehender Graph, welcher wiederverwendet wird.

Anonyme Graphen: Eine Projektion, welche nur für eine Abfrage erzeugt und dann verworfen wird.

Graphen können auf zwei unterschiedliche Arten erzeugt werden:

Native Projektion: Es wird eine Liste von Labeln, Knoten, Beziehungen und Eigenschaften angegeben, welche geladen werden.

Cypher-Projektion: Knoten und Beziehungen werden durch Cypher-Ausdrücke definiert. Diese Projektion ist mächtiger, aber langsamer als eine native Projektion. Empfohlen wird, dass Cypher-Projektionen nur in der Entwicklungsphase angewendet werden, um später die Datenbank so aufbauen zu können, dass eine native Projektion ermöglicht wird.

Für noch höhere Abstraktion können virtuelle Knoten und Beziehungen erstellt werden. Diese existieren nicht innerhalb des Graphen, sondern werden nur als Ergebnis einer Abfrage erzeugt. Virtuelle Einträge sind beispielsweise nützlich, um Cluster zu visualisieren, Informationen zu abstrahieren, Knoten von längeren Pfaden zusammenzufassen oder sicherheitskritische Information zu verbergen. Das Schema ist in Abb. 4.1 abgebildet.

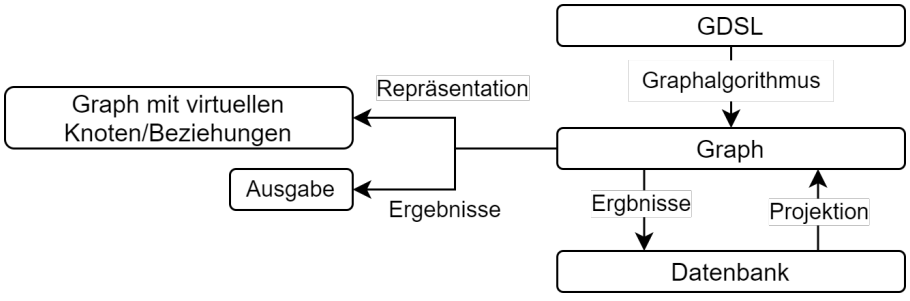


Abbildung 4.1: Die einzelnen Schritte des Workflows mit der Datenbank. Aus der Datenbank werden Graphen erstellt, welche eine Teilmenge der Information der Datenbank enthalten. Auf einen so vorbereiteten Graphen werden die gewünschten Algorithmen angewendet. Die Ergebnisse können je nach Wunsch verworfen, in den Graphen oder in die Datenbank gespeichert werden. (Quelle: eigene Darstellung)

Awesome Procedures On Cypher. Awesome Procedures On Cypher (APOC) ist ein Addon für Neo4j. Es erweitert die Standard-Bibliothek um hunderte Funktionen für die verschiedensten Anwendungszwecke. APOC umfasst Funktionen für:

- Datenimport und -export
- Integration mit anderen Datenbanken
- Datensammlung, Formatierung und Umwandlung
- Sammeln von Metadaten der Datenbank
- Dynamisch erzeugte Knoten und Verbindungen
- Pfadfindungsalgorithmen
- Virtuelle Knoten und Beziehungen

Graph Data Science Library. Die Graph Data Science Library (GDSL) ist eine Bibliothek, welche sich auf Graph Data Science spezialsiert. Sie bietet effizient implementierte Funktionen zur mathematischen Analyse von Knoten und Beziehungen. Diese umfassen:

- Pfadfindungsalgorithmen
- Zentralitätsbestimmung
- Gruppenerkennung
- Ähnlichkeitsbestimmung
- Link-Vorhersage
- Knoten-Embeddings
- Knoten-Klassifizierung

Aufgrund der Berechnungskomplexität der verwendeten Algorithmen setzt die GDSL eine Graphprojektion der Datenbank voraus. Die GDSL bietet vier Ausführungsmodi für die Algorithmen:

Stream: Die Ergebnisse werden als Cypher-Ergebnis in Tabellenform zurückgegeben.

Stats: Ergebnisse werden zu einem statistischen Überblick zusammengefasst. Dieser Überblick enthält das Maximum, Minimum sowie einige Perzentile.

Mutate: Die Ergebnisse werden im In-Memory-Graphen gespeichert, ohne die Datenbank zu verändern. Dies ist sinnvoll, wenn Algorithmen auf dem Ergebnis von vorherigen Algorithmen aufbauen, Ergebnisse von mehreren Algorithmen zusammen dargestellt werden oder Ergebnisse abgefragt werden sollen, ohne die Datenbank zu verändern.

Write: Die Ergebnisse werden in die Datenbank geschrieben.

4.3.3 Interaktion mit dem Datenbanksystem.

Der einfachste Weg mit der Datenbank zu interagieren ist über das integrierte Browser-Tool. Damit können Cypher-Abfragen mit einer grafischen Oberfläche direkt an das

System gesendet werden und Ergebnisse werden als Tabelle oder Text zurück gegeben. Dies dient in erster Linie dazu, während der Entwicklungsphase Abfragen leicht testen und debuggen zu können. Um eine Neo4j-Datenbank von einem externen Programm ansprechen zu können muss auf eine API zugegriffen werden. Dafür existieren zwei Protokolle. Dies sind das Bolt-Protokoll und das Hypertext-Transfer-Protokoll (HTTP). Anhang B enthält zwei Beispiele, wie Cypher in der Praxis aussieht.

Bolt: Bolt ist ein binäres Protokoll, welches Cypher-Typen, TLS-Zertifikate und Authentifizierung unterstützt. Um Bolt zu verwenden, wird ein Treiber benötigt. Offizielle Treiber stehen für die Programmiersprachen .NET, Java, Spring, JavaScrip, Go und Python zur Verfügung. Daneben gibt es eine Reihe inoffizieller Community-Treiber, welche andere Sprachen unterstützen oder abweichende Funktionalität bieten. Der Bolt-Server läuft standardmäßig auf `bolt://localhost:7687`.

HTTP: Neben den offiziellen Treibern bietet Neo4j die Möglichkeit, über HTTP zu kommunizieren. Dabei werden Cypher-Abfragen mittels HTTP-POST-Befehl gesendet. Eine einfache HTTP-Abfrage könnte lauten:

```
1  :POST /db/data/transaction/commit {"statements":[
2    {"statement":"CREATE␣(p:Person␣{firstName:␣$name})␣
       RETURN␣p",
3     "parameters":{"name":"Daniel"}}
4  ]}
```

Listing 4.2: Es wird ein *Person*-Knoten mit dem Vornamen Daniel erstellt und ausgegeben.

Einige der Treiber benutzen HTTP intern und stellen nur ein einfacheres Interface zur Verfügung. Der HTTP-Server läuft auf `http://localhost:7474`.

4.4 Zusammenfassung

Es wurde ein Überblick über die verschiedenen Werkzeuge gegeben, mit welchen die einzelnen Schritte der Graphenanalyse durchführbar sind. Dazu wurde ein einfacher Katalog von Kriterien definiert, nach welchem eine Auswahl verschiedener Werkzeuge bewertet wurde. Die Graphendatenbank Neo4j erreichte eine sehr gute Bewertung und wurde daher als Programm für die Umsetzung ausgewählt. Im zweiten Teil wurde ein Überblick über Konzepte und Funktionen von Neo4j gegeben, welche für die Auswertung verwendet wurden.

5 Beschreibung des Datensatzes

In diesem Kapitel werden die Daten vorgestellt, welche verwendet wurden.

5.1 Rohdaten

Es wurden Twitter-Konversationen von Accounts heruntergeladen, bei denen davon ausgegangen wird, dass sie besonders kontrovers diskutiert werden. Diese umfassen (die Twitter-Handles): `nikitheblogger`, `Karl_Lauterbach`, `c_drosten`, `ArminLaschet`, `Markus_Soeder`, `FriedrichMerz`, `OlafScholz`, `KuehniKev`, `ABaerbock`, `c_lindner`, `SWagenknecht`, `Beatrix_vStorch`, `Alice_Weidel`, `MalteKaufmann`, `georgrestle`, `janboehm`, `TeamKenFM`, `RolandTichy`, `reitschuster`, `attila_hildmann`, `RenaadeS`, `ChakerRonai`, `Bruno54312`, `igorpianist` und `greenpeace_de`. Die gesammelten Tweets stammen aus dem Zeitraum vom 19.04.2021 bis 25.06.2021 (Abb. 5.1).

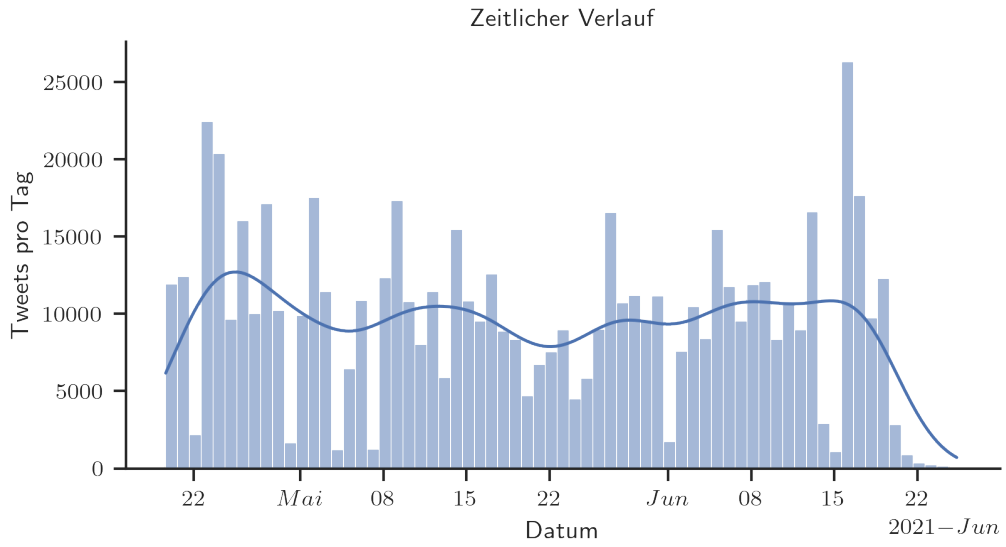


Abbildung 5.1: Zeitliche Verteilung der Aktivität. Jeder Balken zeigt die Anzahl Tweets von einem Tag.

Die Daten wurden von Twitter in einer CSV-Datei heruntergeladen. Diese enthielt einzelne Tweets. Diese CSV-Datei funktioniert wie eine Adjatenzliste, in welcher für jeden Knoten definiert ist, wie er mit anderen Knoten verbunden ist[17]. Die Felder, welche Twitter bietet, sind in Tabelle 5.1 beschrieben. Für alle Nutzer sind die User-ID, der Anzeigenamen, sowie der Twitter-Handle bekannt. Zuletzt wurden die Nutzer markiert, welche gesichert wurden und Toxizitätswerte aller Tweets hinzugefügt.

[17] Alternativ kann ein Tweet auch wie eine Hyperkante interpretiert werden, welche mehrere Nutzer, Hashtags etc. verbindet. In diesem Fall entspricht es einer Inzidenzliste.

Feld	Bedeutung
tweet_id	ID des Tweets
comment_type	main: Der erste Tweet einer Konversation. side: Antworten auf einen Tweet
conversation_id	Bei main-Tweets: Die Tweet-ID Bei side-Tweets: Die Tweet-ID des Main-Tweets, auf den (indirekt) geantwortet wurde
text	Text des Tweets
author_id	Nutzer-ID des Tweet-Autors
ref_type	Bei Side-Tweets: Die Art der Beziehung zum Main-Tweet: replied_to: Direkte Antwort retweeted: Retweet ohne eigenen Kommentar quoted: Retweet mit eigenem Kommentar
ref_id	Bei Side-Tweets: Die Tweet-ID des Tweets, auf den geantwortet wurde.
in_reply_to_user_id	Nutzer-ID des Benutzers, auf den geantwortet wurde
created_at	Erstellugsdatum des Tweets in ISO-Form
mentions	Namen von Nutzern, welche im Text erwähnt wurden
url	Verlinkte URLs
hashtags	Verwendete Hashtags
like_count	Anzahl „Gefällt mir”-Angaben
retweet_count	Anzahl Retweets
quote_count	Anzahl Retweets mit Kommentar
reply_count	Anzahl Antworten

Tabelle 5.1: Beschreibung der Felder des Datensatzes

Aus diesen Feldern wurde die Graphenstruktur extrahiert. Dazu wurden aus der `author_id` *Nutzer*, aus der `Hashtags`-Liste *Hashtags* und aus der `conversation_id` *Konversationen* erstellt. Die `ref_id` verweist auf einen Tweet, `in_reply_to_user_id` auf eine Nutzer-ID und `Mentions` auf Nutzer-Namen (Abb. 5.2). Die genaue Funktionsweise von Konversationen ist in Abb. 5.3 abgebildet.

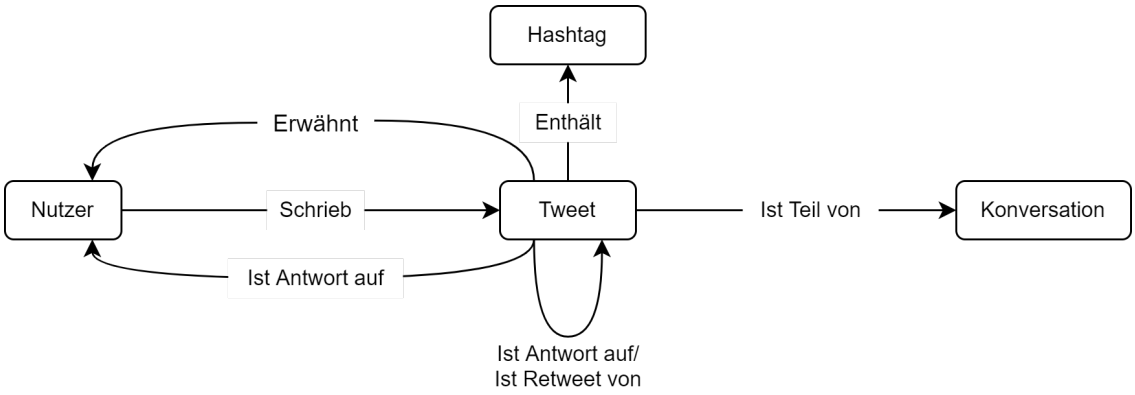


Abbildung 5.2: Alle Knoten-Label und möglichen Verbindungen zwischen diesen Labeln (Quelle: eigene Darstellung).

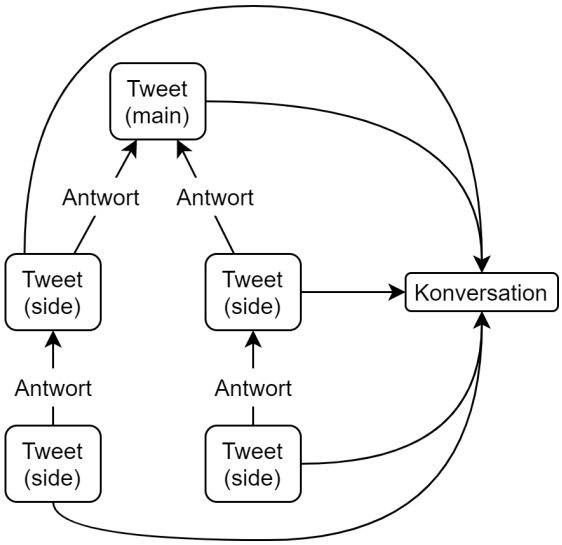


Abbildung 5.3: Die Funktionsweise von Konversationen. Antwort-Beziehungen verbinden einen Antwort-Tweet mit einem Tweet, auf den geantwortet wurde. Konversationen verbinden einen Main-Tweet sowie alle Tweets, die direkt oder indirekt auf diesen Main-Tweet antworteten (Quelle: eigene Darstellung).

5.2 Überblick

Um einen ersten Überblick über den Datensatz zu bekommen wurden einige statistische Kennwerte berechnet. Der Datensatz besteht aus 733.123 Knoten sowie 2.270.210 Verbindungen. Die genaue Zusammensetzung ist in Tabelle 5.2 aufgeschlüsselt.

Knoten	Anzahl
Tweet	623.517
Konversation	4.506
Nutzer	101.214
Hashtag	20.823
Gesamt	733.123

(a) Knoten

Verbindungen	Anzahl
SCHRIEB	603.300
IST_TEIL_VON	607.806
NAHM_TEIL_AN	363.139
IST_ANTWORT_AUF_TWEET	611.919
ENTHAELT	84.046
Gesamt	2.270.210

(b) Verbindungen

Tabelle 5.2: Zusammensetzung der Einträge der Datenbank.

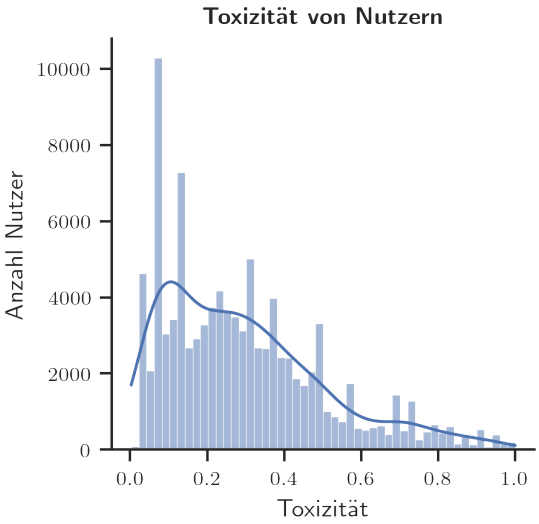
Die Verteilung von identischen Tweets zeigt kaum Auffälligkeiten. Bei gezielten Troll-Angriffen würde ein Akteur unter anderem versuchen identische Texte über viele verschiedene Accounts zu verbreiten.

Alle Tweet-Texte wurden durch die Jigsaw-API (Jigsaw u. a. [o. D.]) von Google in verschiedenen Kategorien bewertet (Tabelle 5.3).

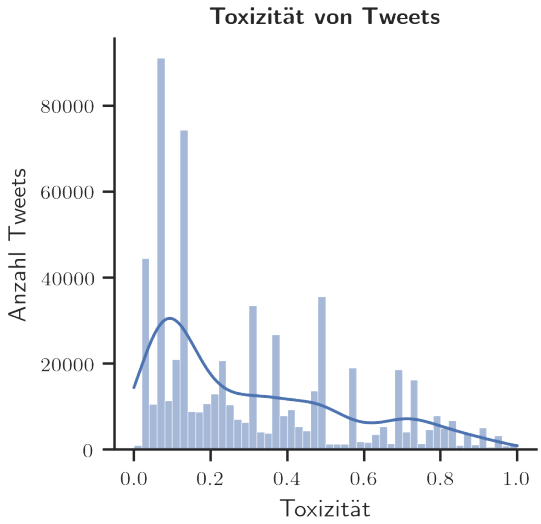
Kategorie	Beschreibung
Toxizität	Ein unanständiger, unvernünftiger oder respektloser Kommentar, durch welchen Personen eine Diskussion verlassen.
Schwere Toxizität	Ein sehr hasserfüllter oder aggressiver Kommentar, durch welchen Nutzer die Diskussion verlassen oder ihren Standpunkt nicht mehr vertreten. Dieses Attribut reagiert viel weniger empfindlich auf mildere Formen der Toxizität, wie z. B. Kommentare, die eine positive Verwendung von Schimpfwörtern enthalten.
Fluch	Schimpfwörter oder andere obszöne oder profane Sprache.
Beleidigung	Beleidigender, aufrührerischer oder negativer Kommentar gegenüber einer Person oder einer Gruppe von Menschen.
Identitätsangriff	Negative oder hasserfüllte Kommentare, welche jemandes Identität angreifen.
Bedrohung	Beschreibt die Absicht, einer Person oder Gruppe Schmerzen, Verletzungen oder Gewalt zuzufügen.

Tabelle 5.3: Beschreibung der Kategorien, welche von Jigwaw bewertet werden. Die Klassifizierung der Trainingsdaten erfolgte binär, also zutreffend oder nicht zutreffend. Jede Bewertung ist im Intervall $[0, 1]$ und beschreibt die Wahrscheinlichkeit, dass ein Nutzer einen Text in der entsprechenden Kategorie einordnen würde.

Wenn die für alle Nutzer die durchschnittliche Toxizität ihrer Tweets betrachtet werden, sind wenige Nutzer besonders toxisch, mit einem Abfall bis hin zu überhaupt nicht toxischen Nutzern (Abb. 5.4).



(a) Verteilung der Toxizität von Nutzern (Durchschnitt aller Toxizitätswerte der Tweets eines Nutzer.).



(b) Verteilung der Toxizität von Tweets

Abbildung 5.4: Die Verteilung der Toxizitätswerte im Datensatz. Der größte Teil der Tweets sowie Nutzer ist wenig bis mittel toxisch, mit einigen wenigen sehr toxischen Einträgen. Die sehr hohen Ausschläge bei einigen Werten der Tweet-Toxizität sind auffällig, können aber an dieser Stelle nicht genauer erklärt werden. Bestimmte Tweet-Eigenschaften scheinen immer gewisse Toxizitätswerte zu erzeugen.

Die Verteilungen der einzelnen Reaktionen sind extrem schief. Wenige Nutzer beziehungsweise Tweets sind sehr dominant, während der größte Teil kaum relevant ist (wenige Reaktionen erhält, Abb. 5.5). Dies stützt die Hypothese, dass durch Betrachtung der Knoten-Werte einige wenige, sehr relevante Knoten bestimmt werden können.

Allerdings reichen diese Informationen allein nicht aus, um Rückschlüsse auf die Toxizität zu ziehen. Die „Gefällt Mir”-Angaben, Retweets, Zitat-Tweets und Antworten korrelieren stark miteinandern. Aus diesem Grund ist es ausreichend, nur die Toxizität weiter zu beachten. Dagegen ist keine Korrelation zwischen diesen Werten und den Toxizitäts-Kategorien zu erkennen (Abb. 5.6).

Die Tweets in Tabelle C.5 zeigen bereits, dass „Hass und Hetze”, „Drohungen, Aufrufe zur Gewalt und Straftaten, menschenverachtende Beleidigungen” und „rassistisch[e] und abstoßend[e]” Äußerungen ein bedeutendes Thema der Konversationen darstellen.

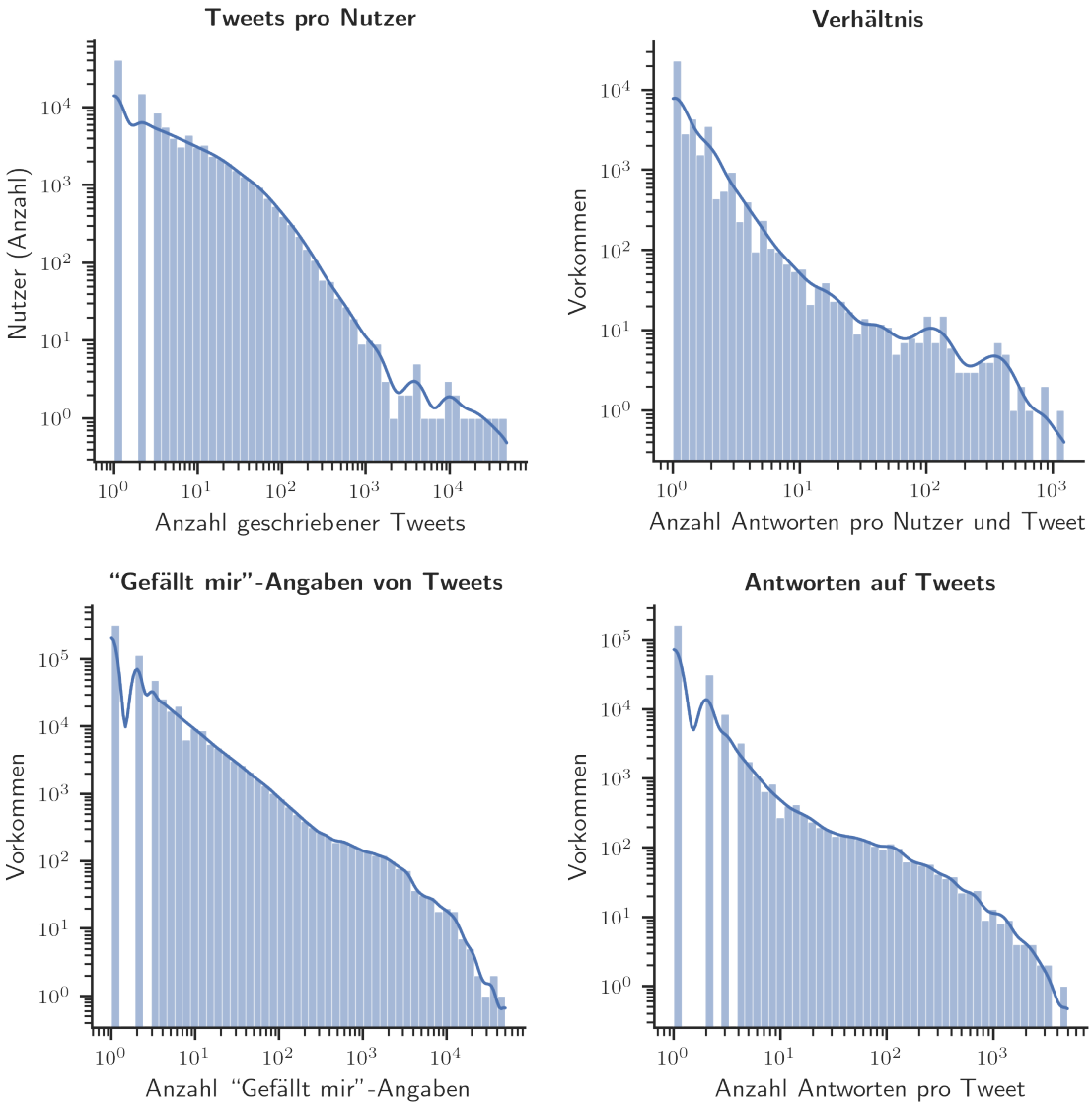


Abbildung 5.5: Die Verteilung von geschriebenen Tweets, erhaltenen Antworten pro Tweet, erhaltenen „Gefällt mir”-Angaben und Antworten pro Tweet. Alle vier Verteilungen werden von einigen wenigen Nutzern dominiert, welche tausende beziehungsweise zehntausende Einträge besitzten. Durch die doppelt logarithmische Darstellung wird deutlich, dass die Werte angenähert einer Pareto-Verteilung folgen.

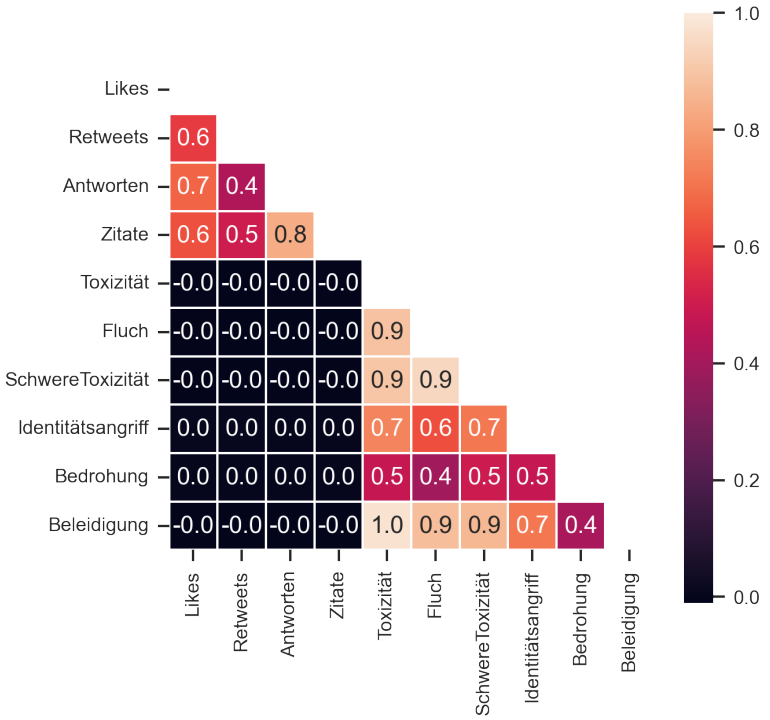


Abbildung 5.6: Pearson-Korrelation vom Logarithmus der „Gefällt mir"-Angaben, Retweets, Zitat-Tweets, Antworten (obere linke Ecke) und Toxizität (untere rechte Ecke) von Tweets. Die Reaktionen sowie die verschiedenen Toxizitäts-Kategorien korrelieren jeweils stark untereinander. Zwischen diesen beiden Merkmalen besteht allerdings keine signifikante Korrelation. Es wurde der Logarithmus der Reaktionen gewählt, da die Korrelation in diesem Fall höher liegt als bei den linearen Werten.

Abbildung 5.7: Schlagwortwolke der häufigsten Hashtags. Die Größe des Begriffs entspricht der Häufigkeit des Vorkommens. Wenn jeder Hashtag ein Thema des Tweets darstellt, lassen sich *Bundestagswahlkampf 2021*, *Covid19-Pandemie*, sowie *Deutsche Politik (allgemein)* als häufigste Themen erkennen.

5.3 Zusammenfassung

In diesem Kapitel wurde der Datensatz vorgestellt. Es wurden die metadaten-Felder der Tweets des Datensatzes beschrieben und erklärt, wie diese zusammenhängen. Anschließend wurden statistische Werte wie die Verteilung von Antworten und „Gefällt mir”-Angaben von Tweets berechnet. Dabei wurde gezeigt, dass diese einer Pareto-Verteilung folgen. Außerdem wurden die Kategorien von Hate Speech erklärt, nach welchen die Tweet-Texte von der Jigsaw-API bewertet wurden. Die Verteilung der Werte weist ein Maximum bei circa $0,3$ auf und fällt bis 1 ab. Die Hate Speech-Kategorien korrelieren sehr stark untereinander, allerdings besteht kein Zusammenhang zu den Reaktionen.

Teil II

Umsetzung und Ergebnisse

6 Wichtungsschema und Projektion

In diesem Kapitel werden die angewendeten Methoden für die Gewichtung und Projektion der Daten erklärt.

Da sich die Auswertungen dieser Arbeit auf die Nutzer beziehen sollen, war es nötig, eine Beziehung zwischen den einzelnen Nutzern zu definieren. Diese Definition ist nicht trivial und hat entscheidenden Einfluss auf die Ergebnisse der weiteren Analysen. Nach Skiena (2020, S. 19, S. 276) ist der mit Abstand wichtigste Schritt des Algorithmus-Entwurfes, die Datenstruktur der Anwendung so zu konzipieren, dass die gewünschte Funktion möglichst mit Standard-Algorithmen erreicht werden kann. Es wird explizit davon abgeraten, neue (Graphen-)Algorithmen zu entwerfen. Daher werden stattdessen einige mögliche Graphprojektionen mit geringen Anpassungen an die Zentralitätsalgorithmen verglichen. Zudem wurde sich daran orientiert, welche Algorithmen die GDSL in Neo4j unterstützt, um den zusätzlichen Programmieraufwand zu minimieren.

Es ist dabei immer zu beachten, dass ein komlexes System niemals eine einzige *richtige* Darstellung hat. Jede Projektion und jede Wichtung kann nur einen Teil aller vorhandenen Informationen betrachten. Daher muss immer zwischen dem Knoten im Netzwerk und der Entität, welche der Knoten repräsentiert, unterschieden werden (Zweig 2016, S. 110, S. 112). Anstatt *„Karl Lauterbach ist der einflussreichste Nutzer."* müsste es heißen *„Der Knoten, welcher in diesem Datensatz Karl Lauterbach darstellt, hat in diesem Modell und mit diesem Wichtungsschema den höchsten Einfluss im Sinne des ArticleRanks."*. Diese Unterscheidung wird zur Vereinfachung im weiteren Verlauf nicht mehr explizit erwähnt.

Zweig (2016, S. 139) zeigt einige Beispiele für Co-Autor-Graphen von wissenschaftlichen Artikeln. Beziehungen in Sozialen Online-Netzwerken können ähnlich modelliert werden. Autoren von Artikeln werden über die Anzahl ihrer gemeinsamen Paper verbunden. Diese Gewichtung wurde anschließend normalisiert. Als einziges Beispiel für die Anwendung speziell in Sozialen Online-Netzwerken wurde die Studie von Allen u. a. (2019) gefunden, welche dieses Thema aufgreift. Kitajima u. a. (2021) verwendeten ein Wichtungsschema, in welchem die Verbindung zwischen einem Autor und einem Post über die Bewertung des Posts gewichtet wurde. Ebenfalls wurde der Jaccard-Index als Ähnlichkeitsmaß verwendet.

In dieser Arbeit wurden drei Modelle implementiert, welche im Folgenden vorgestellt werden. Zuerst wurde eine Funktion entwickelt, um den Einfluss eines einzelnen Tweets anhand seiner ausgelösten Reaktionen zu bestimmen. Dies wurde für die Berechnung der Gewichtung der Beziehungen verwendet.

Im ersten Modell wurde nur betrachtet, wie oft ein Nutzer auf andere Nutzer antwortete.

Die Anzahl Antworten dient als Gewicht der Beziehung. Im zweiten Modell wurden Nutzer und Konversationen betrachtet. Die Beziehung zwischen zwei Nutzern wurde über gemeinsame Teilnahme an Konversationen definiert. Auch hier wurden die Reaktionen auf Tweets als Gewicht der Beziehungen gewertet. Im dritten Modell wurden die Beziehungen der Tweets untereinander direkt betrachtet. Das Ranking der Nutzer wird über die Bewertung der geschriebenen Tweets errechnet.

Die Ähnlichkeit über gemeinsame Konversationen verbindet sehr viele Nutzer miteinander, von welchen viele überhaupt nicht direkt interagierten. Eine Konversation kann aber als ein Thema betrachtet werden, an welchem alle Nutzer bewusst teilnahmen. Die Beziehung über Antworten verbindet nur Nutzer, welche direkt miteinander interagierten. Dafür geht die Information verloren, welche Nutzer indirekt gemeinsam aktiv waren. Eine Modellierung mittels einzelner Tweets sollte die höchste Genauigkeit erreichen, da der Einfluss für jeden Tweet und jede Konversation einzeln bestimmt wird.

Weitere Modelle sind beispielsweise durch den Vergleich von Retweets (vgl. Allen u. a. 2019, S. 386) oder verwendeten Hashtags (Guarino u. a. 2020) möglich. Eine Gewichtung von Konversationen sowie Tweets in Konversationen wurde aufgrund der Komplexität wieder verworfen.

6.1 Wichtung von Tweets

Zuerst wurde für jeden Tweet der Einfluss bestimmt.

Idee 1: Reaktionen. Gemäß Definition 3.1 können die Reaktionen auf einen Tweet mit der Anzahl *„Gefällt mir"-Angaben* gm, *Retweets* rt, *Zitat-Tweets* zt und *Antworten* ant gemessen werden. Jeder Tweet T hat ein Gewicht $\mathbf{E}(T)$:

$$\mathbf{E}(T) = (gm,\ rt,\ zt,\ ant)^{\mathsf{T}}$$

Aufgrund der in Abb. 5.6 gezeigten starken Korrelation der Werte werden diese vier Dimensionen zu einem Wert zusammenfassen:

$$\begin{aligned} E(T) &= \frac{\log(\mathbf{E}(T)+\mathbf{1})}{\#T_N} \\ &= \frac{\log(gm+1)+\log(rt+1)+\log(zt+1)+\log(ant+1)}{\#T_N} \end{aligned}$$

wo T_N die Tweets des Nutzers N darstellen. Die Addition von 1 erfolgt, da ein Großteil der Tweets null Reaktionen erhielt und damit der Logarithmus undefiniert wäre.

Idee 2: Zusätzliche Informationen. Zusätzlich zu den Reaktionen können weitere Informationen betrachtet werden, wie die thematische Analyse des Inhalts (beispielsweise enthaltene Hashtags oder Text-Klassifizierung), Sentimentanalysen (Toxizitätswerte) oder Klassenzugehörigkeiten (*Troll/kein Troll* oder *Strafrechtlich relevant/nicht strafrechtlich relevant*).

Weitere Klasseneinteilungen standen nicht zur Verfügung. Informationen zum Sentiment des Textes wurden nicht mit betrachtet. Stattdessen sollte festgestellt werden, ob ein Zusammenhang zwischen Sentiment und den Reaktionen besteht.

6.2 Projektion

6.2.1 Modell 1: Antworten

Für das erste Modell wurden die Antworten eines Nutzer auf Tweets von anderen Nutzern betrachtet.

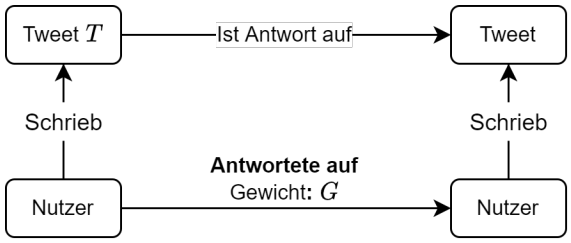


Abbildung 6.1: Nutzer-Nutzer-Beziehung über Antworten auf Tweets.

Als Gewicht der Beziehung werden die Anzahl der Tweets $G_1 = \#T$ beziehungsweise die Summe der Gewichte der einzelnen Tweets $G_2 = \sum_T E(T)$ verwendet.

6.2.2 Modell 2: Konversationen

Für das zweite Modell wurde betrachtet, welcher Nutzer an welchen Konversationen teilnahm. Anschließend wurde dieser Graph mit Nutzern und Konversationen genutzt, um mittels der Jaccard-Ähnlichkeit die Beziehung zwischen einzelnen Nutzern zu berechnen (Abb. 6.3). Es wurden nur Nutzer einbezogen, welche an mindestens 10 Konversationen teilnahmen (8098 Nutzer insgesamt).

Jeder Nutzer hat einen Einfluss auf eine Konversation K, welcher von dem Einfluss seiner Tweets T abhängt (Abb. 6.2): $G = \sum_{T \in K} E(T)$.

Mit diesem Einfluss wird die Jaccard-Ähnlichkeit der Nutzer berechnet (Abb. 6.3).

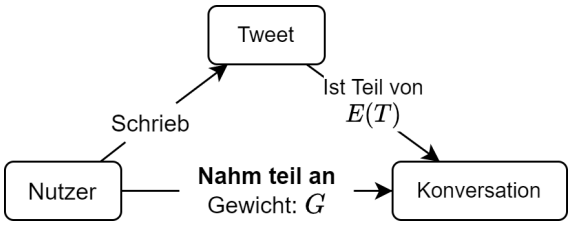


Abbildung 6.2: Einfluss eines Nutzers auf eine Konversation

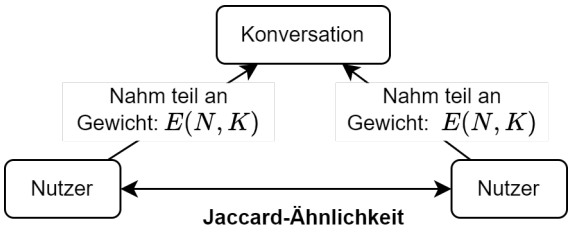


Abbildung 6.3: Projektion des Graphen, welche nur Nutzer und Konversationen enthält. Zwischen zwei Nutzern wurde die Jaccard-Ähnlichkeit bestimmt. Der Einfluss G wird als Gewicht verwendet.

Idee 1 Der Einfluss E eines Benutzers N auf eine Konversation k ist die Summe der Einflüsse aller seiner Tweets T_N in dieser Konversation:

$$E(N,k) = \sum_{T_N \in k} E(T_N)$$

Idee 2 Um den Einfluss eines einzelnen Nutzers nach oben zu beschränken kann es sinnvoll sein, nicht die Summe direkt zu verwenden. Stattdessen kann der Logarithmus, Tangens hyperbolicus oder Okapi25-Wert des Einflusses genutzt werden.

Im Gegensatz zu Spranger u. a. (2020) wird der Einfluss eines Nutzers somit nicht global bestimmt, sondern lokal für jede Konversation einzeln, an welcher der Nutzer teilgenommen hat. Aus diesem Graph wird der globale Einfluss bestimmt.

6.2.3 Modell 3: Tweets

Zuletzt wurden einzelne Tweets betrachtet. Damit werden Beziehungen zwischen Nutzern sowie zwischen verschiedenen Konversationen komplett ignoriert. Als Gewicht der Beziehung wurde der Einfluss des Antwort-Tweets verwendet: $G = E(T)$. Dadurch können

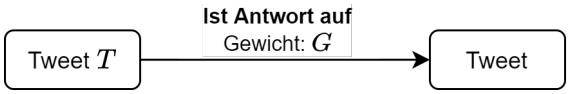


Abbildung 6.4: Das Modell der Beziehungen der Tweets.

einzelne Tweets besser bewertet werden. Allerdings gehen alle Informationen darüber

verloren, welche Nutzer in unterschiedlichen Konversationen miteinander interagierten.

6.3 Zusammenfassung

In diesem Kaptitel wurden die Konzepte erläutert, nach welchen eine sinnvolle Projektion der gegebenen Daten definiert werden kann und welche Probleme dabei zu beachten sind. Dazu wurde zuerst definiert, wie sich der Einfluss eines einzelnen Tweets berechnet. Dieser Einfluss ist einerseits umso größer, je mehr Reaktionen der Tweet erhielt. Andererseits ist der Einfluss kleiner, je mehr Tweets der Nutzer insgesamt schrieb. Damit soll starke Aktivität der Nutzer bestraft werden. Davon ausgehend wurden drei Modelle erarbeitet, welche jeweils verschiedene Aspekte der Daten repräsentieren sollen.

7 Analyse und Visualisierung

Es wurden verschiedene Auswertungen mit den Modellen durchgeführt. Dazu werden die einflussreichsten Nutzer durch die Berechnung von des ArticleRanks bestimmt sowie mit dem Louvain-Algorithmus Gruppen bestimmt. Zuerst erfolgt ein Vergleich zwischen dem PageRank und ArticleRank. Für jedes Modell wird die Verteilung des ArticleRanks und der Gruppengrößen bestimmt. Beim ArticleRank wird untersucht, durch welche Beziehungen dieser Wert zustande kommt. Die einflussreichsten Nutzer werden in Kategorien eingeordnet:

Kategorie	Beschreibung
Politiker	Personen, welche ein poltisches Amt oder eine Partei in ihrer Profilbeschreibung ausweisen.
Journalist	Personen, welche für Zeitungen oder Nachrichtensendungen arbeiten.
Person des öffentlichen Lebens	Sonstige bekannte Personen (Schauspieler, Musiker, Sportler, etc.)
Privatperson	Personen, welche mit einem Klarnamen angemeldet sind, aber keiner der obigen Kategorien zugeordnet werden können.
Partei	Parteipolitischer Account, welcher keinem einzelnen Politiker zuzuordnen ist.
Organisation	Sonstige Accounts, welche einer Gruppe aus mehrern Personen zugeordnet sind.
Unbekannt	Accounts, welche keiner der obigen Kategorien zugeordnet werden können, meist ohne Klarnamen.

Tabelle 7.1: Beschreibung der Nutzer-Kategorien

Im folgenden wird mit *ungewichtet* der Referenzwert bezeichnet, wenn die Anzahl der Antworten beziehungsweise Konversationen zwischen zwei Nutzern als Gewicht der Verbindung betrachtet werden. Dagegen bezeichnet *gewichtet* das Wichtungsschema, welches sich aus den Reaktionen berechnet.

7.1 Modell 1: Antworten

7.1.1 Vergleich von PageRank und ArticleRank

Zuerst wurden der ArticleRank sowie der PageRank bestimmt. Dabei wurden als Gewicht der Kanten nur die Anzahl der Antworten zwischen zwei Nutzern und das durch das Wichtungsschema berechnete Gewicht verwendet. Diese sind im Modell 1 fast gleichwertig (Abb. 7.1). Dies deutet darauf hin, dass bei hinreichender Größe des Netzwerkes die Bedeutung jeder einzelnen Verbindung so gering ist, dass eine unterschiedliche Gewichung keinen Unterschied mehr macht, oder dass die Unterschiede sich gegenseitig aufheben. Zwischen den 1000 einflussreichsten Nutzern besteht eine sehr große Überschneidung (Abb. 7.2).

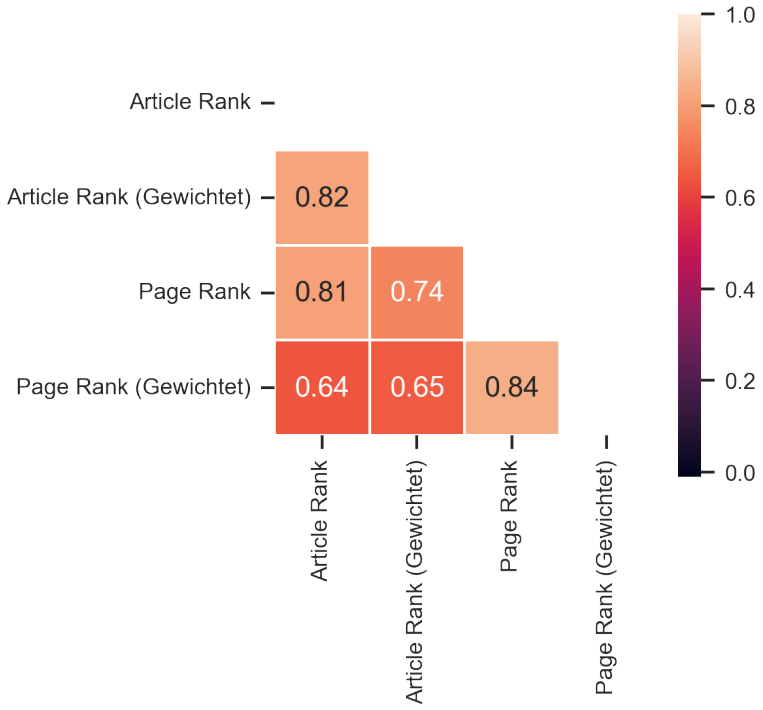


Abbildung 7.1: Spearman-Korrelation der vier Rankings. Alle vier Werte korrelieren sehr stark miteinander. Zwischen dem ArticleRank und dem PageRank ist die Korrelation besonders stark. Es wurde speziell die Spearman-Korrelation gewählt, da bei Zentralitätmaßen meist nicht der numerische Wert interessant ist, sondern nur die Reihenfolge der Knoten. Die vier Ergebnisse korrelieren sehr stark. Der verwendete Algorithmus und das genaue Wichtungsschema haben nur einen sehr geringen Einfluss.

7.1.2 Globale Ergebnisse

Die zehn Top-Nutzer sind:

1. Jens Spahn (CDU)

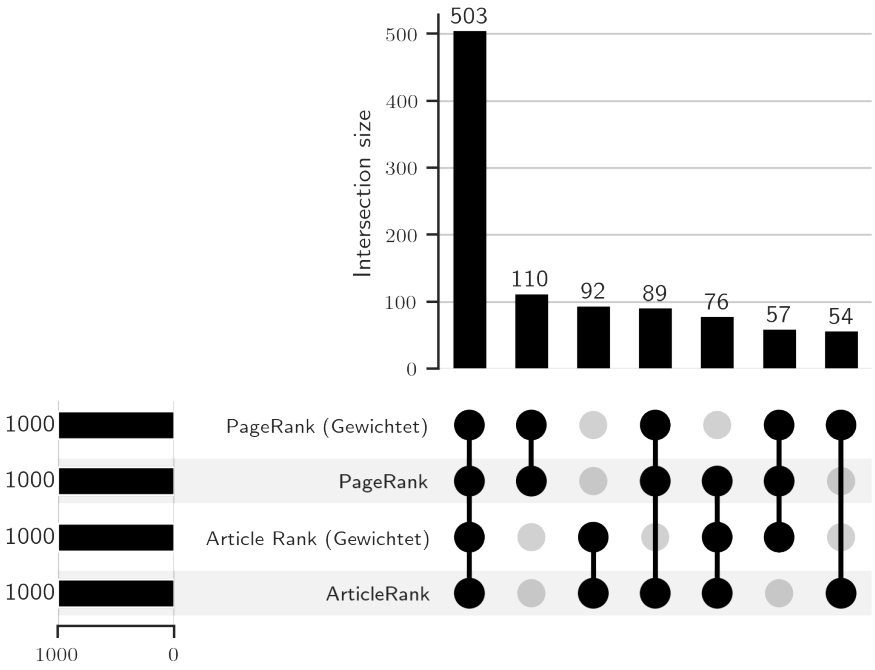


Abbildung 7.2: Die Schnittmengen der 1000 einflussreichsten Nutzer nach Ranking-Algorithmus. 503 Nutzer sind in allen vier Listen vorhanden. 110 Nutzer werden nur vom PageRank, 92 Nutzer nur vom ArticleRank und 222 Nutzer von drei der Algorithmen erkannt. Abgebildet sind nur Schnittmengen aus mindestens zwei Grundmengen und mit mindestens 50 Elementen.

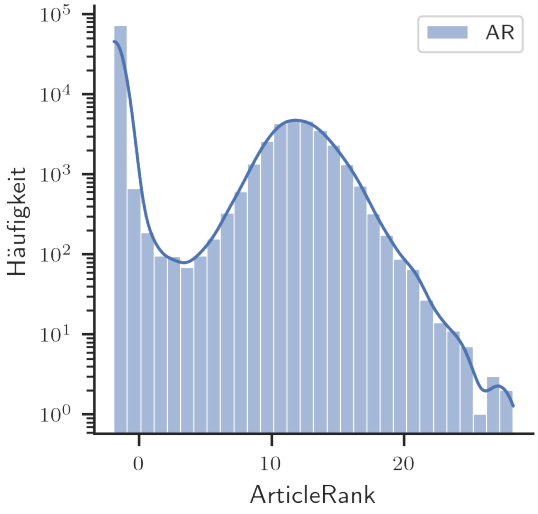


Abbildung 7.3: Verteilung des ArticleRanks der Nutzer.

2. Michael Kellner
3. Manuela Schwesig (SPD)
4. Svenja Schulze (SPD)
5. Lutz van der Horst
6. Micky Beisenherz
7. Philip Plickert
8. Hans-Georg Maaßen (CDU)
9. AfD_Muenster (AfD)
10. Franziska Giffey (SPD)

Im nächsten Schritt wurden die 50 einflussreichsten Nutzer betrachtet und manuell in verschiedene Kategorien eingeordnet (Abb. 7.4)[18]. Spranger u. a. (2020) argumentieren, dass Meinungsführer in einem Netzwerk eher Individuen (Journalisten, Politiker) als Medien-Accouts oder Parteien sind. Dieser Argumentation folgend ist der ungewichtete ArticleRank am besten geeignet, Meinungsführer auszumachen. Die Ergebnisse sind bei allen vier Variationen sehr ähnlich.

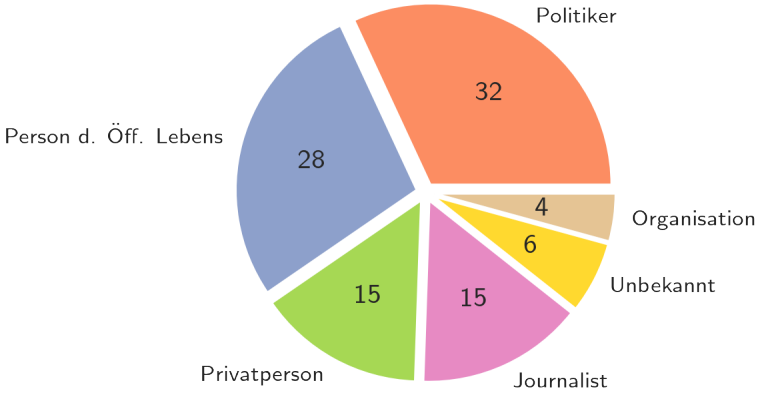


Abbildung 7.4: Die Kategorien der Top 50 durch den ArticleRank (mit Wichtungsschema) als einflussreich erkannten Nutzer (Annotation in Prozent). Es wurden nur Nutzer betrachtet, welche nicht aktiv gescrapted wurden.

Es konnte kein relevanter Zusammenhang zwischen dem ArticleRank eines Nutzers und der Toxizität seiner Tweets festgestellt werden ($r = 0,01$, $p = 0,09$). Bei der Bestimmung von *Meinungsführern* exisierte keine Vorgabe, bestimme Nutzer zu bewerten, sondern der Datensatz solle explorativ untersucht werden. Daher ist es sinnvoll, die Auswertungen auf die einflussreichsten Nutzer zu beschränken. Durch Testen verschiedener Werte wurde eine Anzahl von 1000 Nutzern festgelegt. Diese Zahl ist dabei so gewählt, da die untersuchten Zusammenhänge messbar stärker sind, und statistische Auswertungen ein hohes Signifikanzniveau haben.

Es besteht eine Korrelation, wenn nur die relevantesten Nutzer betrachtet werden

[18] Eine umfangreichere Auswertung war im Rahmen dieser Arbeit nicht möglich.

(Abb. 7.5 und 7.6). Die einflussreichsten Nutzer sind im Durchschnitt signifikant weniger toxisch als der Rest der Nutzer.

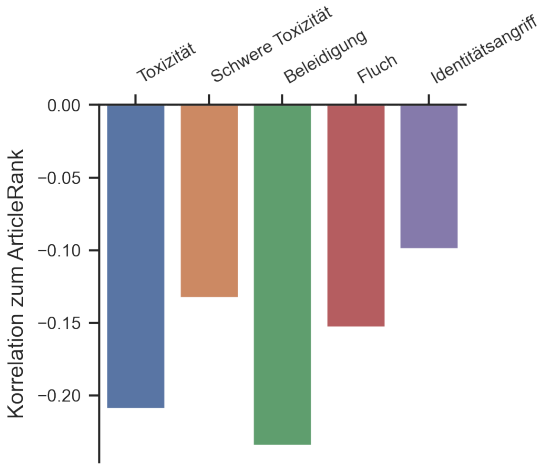


Abbildung 7.5: Die Tabelle zeigt die Korrelation zwischen dem ArticleRank der 1000 einflussreichsten Nutzer und den durchschnittlichen Toxizitätswerten ihrer Tweets. Die stärkste negative Korrelation besteht zu *Toxizität* und *Beleidigung*.

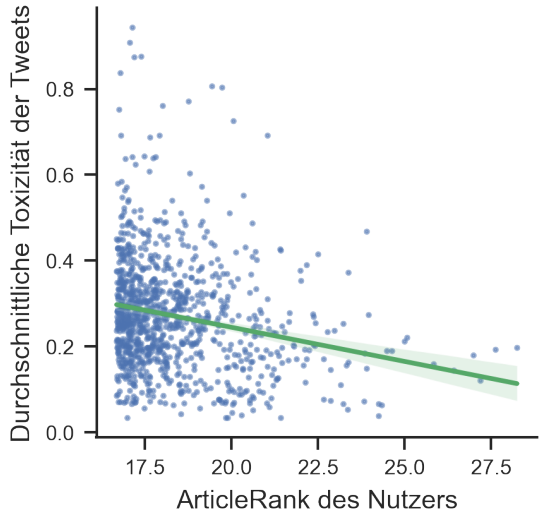


Abbildung 7.6: Zusammenhang zwischen Toxizität und ArticleRank der 1000 einflussreichsten Nutzer. Der Korrelationskoeefiziert beträgt $r = -0,20$ und das Signifikanzniveau $p = 2,4 * 10^{-11}$.

Diese Korrelation kann auf zwei Arten erklärt werden. Einerseits, da *Toxizität* bei Jigsaw darüber definiert wird, dass der Kommentar dazu führt, dass Nutzer die Konversation verlassen. Daher erhalten toxische Nutzer weniger Antworten, was wiederum zu einem geringeren ArticleRank führt. Andererseits kann davon ausgegangen werden, dass Personen des öffentlichen Lebens (welche oft einflussreich sind) bewusst weniger toxisch schreiben, da dies negative Auswirkungen auf die öffentliche Wahrnehmung hat oder sogar strafrechtlich verfolgt werden kann.

Der PageRank sowie der ArtickeRank eignen sich, Meinungsführer zu finden. Allerdings besteht ein starker Bias, welcher die Nutzer bevorzugt, deren Konversationen überwacht wurden. Dies wird verstärkt, da durch den gerichteten Charakter der Projektion der PageRank nur in eine Richtung verbreitet werden kann.

7.1.3 Erhaltene und geschriebene Antworten

Um die Position der Nutzer im Netzwerk genauer zu beschreiben, wurden die Nachbarn aller Nutzer bestimmt (Abb. 7.7). Die Verteilung der Werte wird in Abb. 7.8 dargestellt. Nutzer mit einem ArticleRank kleiner als 10 können als Inaktive angesehen werden, da sie kaum aktiv sind und kaum Antworten von anderen Nutzern bekommen.

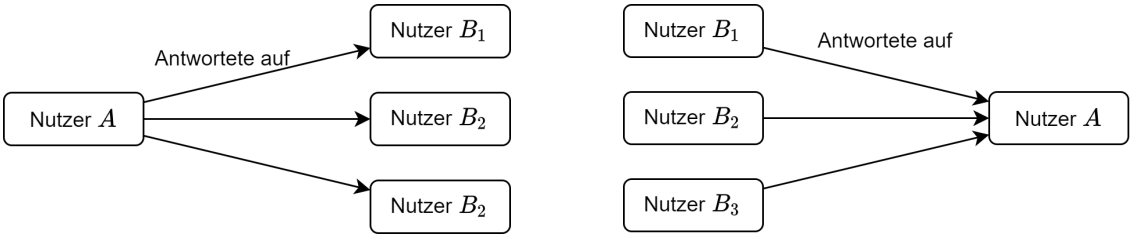


Abbildung 7.7: Für jeden Nutzer A wurde der durchschnittliche ArticleRank
(a) aller ausgehenden Verbindungen sowie
(b) aller eingehenden Verbindungen B_i bestimmt.

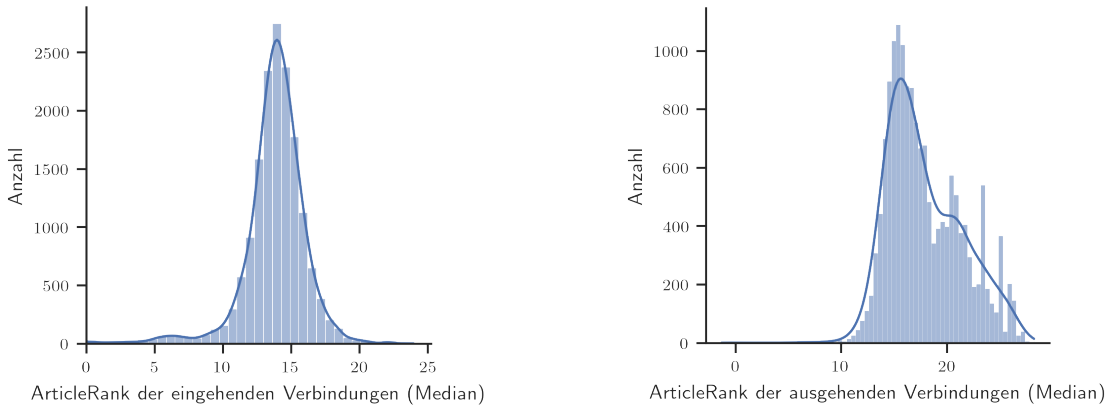
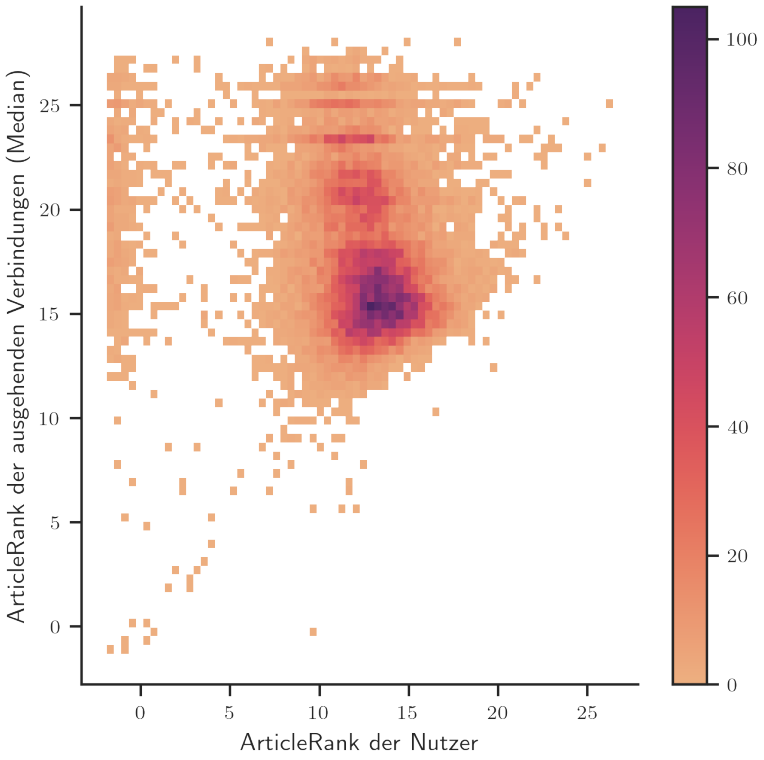
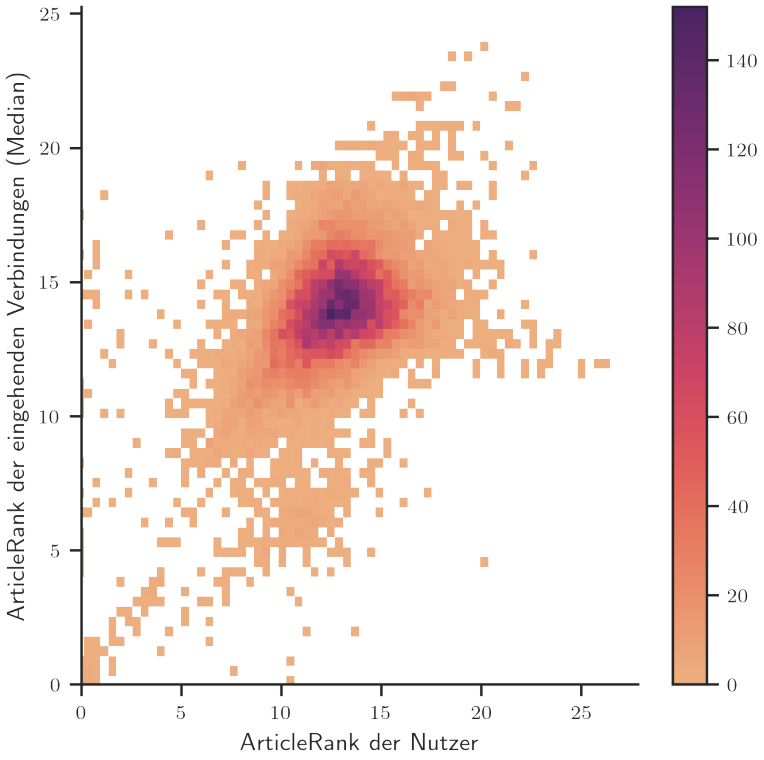


Abbildung 7.8: (a) Die meisten Nutzer erhalten Antworten von durchschnittlich einflussreichen Nutzern (ArticleRank von circa 14). Fast kein Nutzer erhält nur Antworten von sehr einflussreichen Nutzern (ArticleRank über 20).
(b) Die aktivsten Nutzer sind Nutzer mit einem ArticleRank von 15, mit einem Abfall zu 10 und 25 hin.

Dabei gibt es eine klare Ordnung der Nutzer (Abb. 7.9). Antworten geschenen fast immer vom weniger einflussreichen Nutzer hin zum einflussreicheren Nutzer. Interaktionen zwischen den stärksten Meinungsführern finden kaum statt.



(a)



(b)

Abbildung 7.9: (a) Fast alle Nutzer antworten nur Nutzern mit einem ähnlichen oder höheren ArticleRank als ihr eigener. Durch die kleine Bin-Größe wird sichtbar, dass sich die Antworten bei Meinungsführern auf einige weniger Nutzer konzentrieren (horizontale Linien).
(b) Fast alle Nutzer bekommen Antworten von Nutzern, welche einen ähnlichen ArticleRank wie sie selbst haben. Auffällig ist dass die Meinungsführer mit einem ArticleRank > 20 nur Antworten von deutlich weniger einflussreichen Nutzern bekommen.

Das legt eine leicht abgeänderte Variante von Definition 3.1 nahe. Als *Meinungsführer* kann man auch die Nutzer betrachten, welche die meisten Antworten von anderen einflussreichen Personen erhalten. Diese sind:

1. Miguel Robitzky
2. Marie von den Benken
3. Alman-Nick
4. Christina Schlag
5. Dominik Scharpf
6. (Evil B) JoernCarmaker
7. Titanix
8. Jana Frielinghaus
9. Elle Kotzo
10. Marie-Agnes Strack-Zimmermann (FDP)

Während die in Abschnitt 7.1.2 genannten Accounts meist Bundespolitiker oder bekannte Fernsehpersönlichkeiten sind, sind die Personen in dieser Liste Lokalpolitiker, Zeitungsredakteure und Personen, welche nur auf Twitter bekannt sind.

7.1.4 Auswertungen nach Gruppen

Für eine differenzierte Auswertung wurden mit dem Louvain-Algorithmus Gruppen im gewichteten Graphen bestimmt (Abb. 7.10). Anschließend wurde für jede der fünf größten Gruppen der ArticleRank nur zwischen den Gruppenmitgliedern berechnet. Diese Ergebnisse sind bereits deutlich aussagekräftiger. Zu jeder Gruppe wurden die einflussreichsten Nutzer stichprobenartig analysiert und grob eingeordnet[19].

Die größte Gruppe enthält Politiker wie Aminata Touré (Bündnis 90/Die Grünen), Johannes Vogel (FDP), Saskia Esken und Nancy Faeser (beide SPD). Des Weiteren sind die Fernseh-Moderatoren Micky Beisenherz und Fabian Köster enthalten. Die letzte Kategorie lässt sich grob als *Corona-Leugner* beziehungsweise *Corona-Maßnahmen-Kritiker* beschreiben (Argo Nerd, Hannovergenuss, ZeroCool1002, etc.). Diese Nutzer teilen oft auch konservative bis rechte Inhalte.

Die zweitgrößte Gruppe dreht sich um Autoren wie Jan Böhmermann, Miguel Robitzky, Markus Hennig und Klaas Heufer-Umlauf. Daneben sind die Journalistin Teresa Eder, die Politiker Marco Buschmann (FDP) und Norbert Röttgen (CDU) und der Meteorologe Sebastian Keßler eingeordnet.

Die dritte Gruppe besteht aus Annalena Baerbock (Kanzlerkandidatin für Bündnis 90/Die Grünen), gefolgt von vielen liberal bis konservativ/rechten Journalisten und Politikern (Christian Lindner, Roland Tichy, Marcus Pretzell, Anna Dobler).

[19] Eine umfassende Auswertung der einzelnen Gruppen war im Rahmen dieser Arbeit nicht möglich.

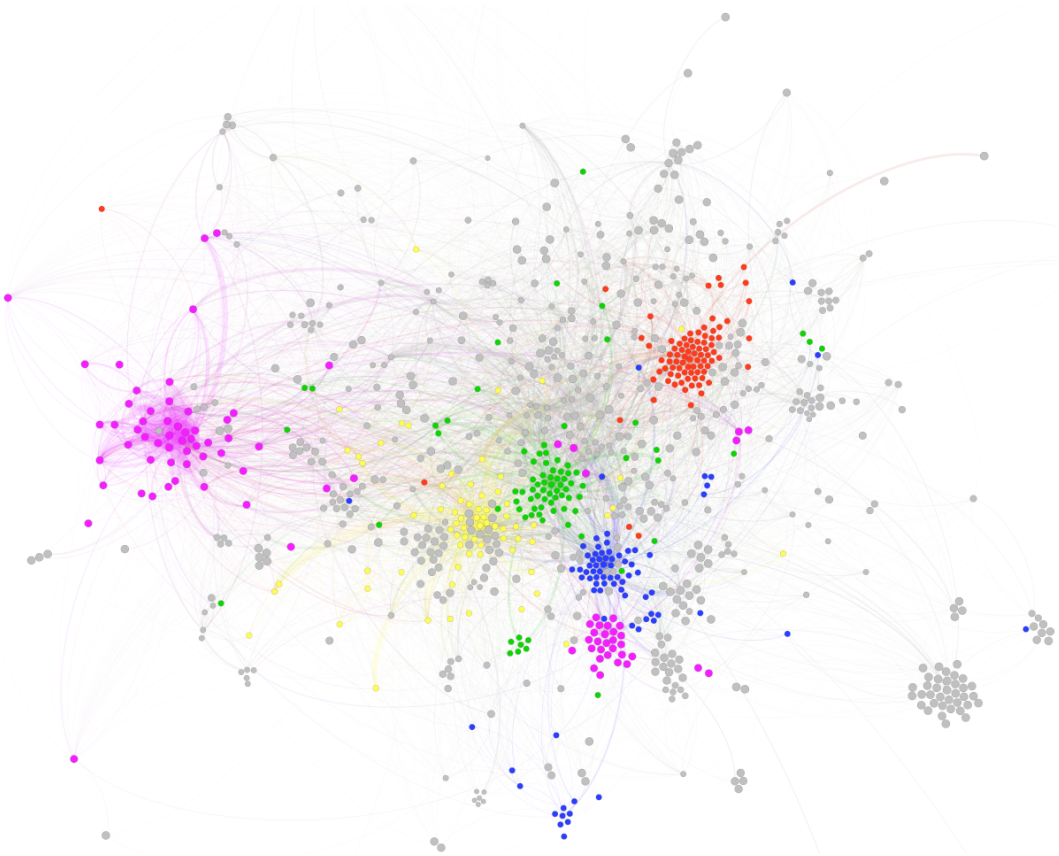


Abbildung 7.10: Darstellung des Graphen der einflussreichsten Nutzer. Die Größe des Knoten stellt den PageRank und die Farbe die Gruppenzugehörigkeit (der größten Gruppen) dar. Die Dicke der Verbindung stellt die Stärke dar. Gruppe 1: Pink, Gruppe 2: grün, Gruppe 3: rot, Gruppe 4: blau, Gruppe 5: gelb.

Die vierte Gruppe enthält die Musiker Igor Levit und Julia Hagen, die Autorin Carolin Emcke, den Journalisten Peter Ahrens und die Autorin und Journalistin Marie von den Benken.

Die fünfte Gruppe enthält Karl Lauterbach (SPD) und Alice Weidel (AfD). Sie enthält keine weiteren bekannten Personen, sondern hauptsächlich „Corona-Leugner" und 'Corona-Maßnahmen-Kritiker", ähnlich wie Gruppe 1. Daher wurden auch keine kleineren Gruppen mehr betrachtet.

Auffällig ist, dass sowohl inhaltlich sehr ähnliche als auch sehr gegensätzliche Nutzer gemeinsam eingruppiert wurden. Insbesondere Annalena Baerbock und Karl Lauterbach wurden in Gruppen eingeordnet, welche dem Gegenstück ihrer politischen Ausrichtung entsprechen. Dies kann damit erklärt werden, dass sie besonders stark von politischen Gegnern und „Trollen"[20] kommentiert werden. Zudem entspricht es dem thematischen Überblick aus Abb. 5.7.

In jeder Gruppe sammeln sich viele Nutzer mit geringerem ArticleRank um einige wenige Nutzer mit höherem ArticleRank. Dies lässt darauf schließen, dass jede Gruppe eigene Meinungsführer besitzt. Da diese Gruppen oft inhaltlich sehr gegensätzlich aufgebaut sind, kann nicht davon ausgegangen werden, dass die Gruppen einzelne Filterblasen darstellen. Viel mehr lässt sich sagen, dass Nutzer aus unterschiedlichen (realen) Gruppen auf Twitter sehr stark miteinander interagieren.

7.2 Modell 2: Konversationen

Das Modell der Konversationen scheint schlechter zu sein, das Netzwerk sinnvoll darzustellen. Jede Konversation verbindet alle teilnehmenden Nutzer zu einem vollverknüpften Teilgraphen. Die erkannten Gruppen bestehen daher aus Nutzern, welche an Konversationen mit besonders vielen anderen Teilnehmern teilnahmen. Anstatt einer hierarchischen Verteilung bestehen die Gruppen aus vielen Nutzern mit ähnlichem ArticleRank. Daher lassen sich einzelne Meinungsführer kaum ausmachen. Die zehn Top-Nutzer sind:

1. Steffen Bender
2. volkspozilei
3. Nikolas Weber
4. bombiiii
5. Herr W.
6. Roland Wissel
7. Fanboy of Disruption
8. Fred_Groeger
9. MZuhlsdorf

[20] Eine genaue Definition eines *Trolls* war in dieser Arbeit nicht möglich, daher kann der Begriff auch nur sehr bedingt verwendet werden.

10. wolliane

Alle diese Nutzer gehören zu den Kategorien *Unbekannt* oder *Privatperson*. Die Gruppen erlauben keine brauchbare Auswertung. Allerdings sind die Zugehörigkeiten ähnlich zu denen im Antworten-Modell (Abb. 7.11). Die Verteilung des ArticleRanks unterscheidet sich stark vom Antworten-Modell (Abb. 7.12). Insgesamt sind die meisten Nutzer duetlich weniger einflussreich und bei der Betrachtung der Beziehungen werden weniger Muster sichtbar.

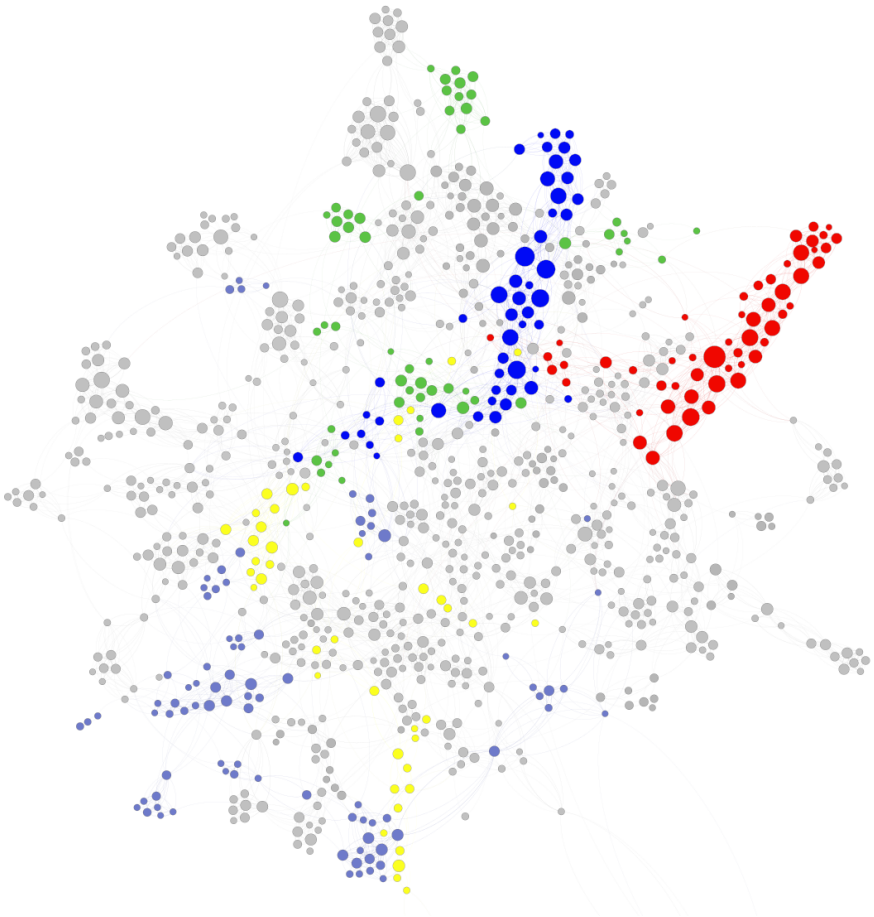


Abbildung 7.11: Darstellung des Graphen der Konversationen. Auch hier stellt die Größe des Knoten den PageRank und die Farbe die Gruppenzugehörigkeit dar.

In diesem Modell besteht kein signifikanter Zusammenhang zwischen Toxizität und Einfluss (Abb. 7.13).

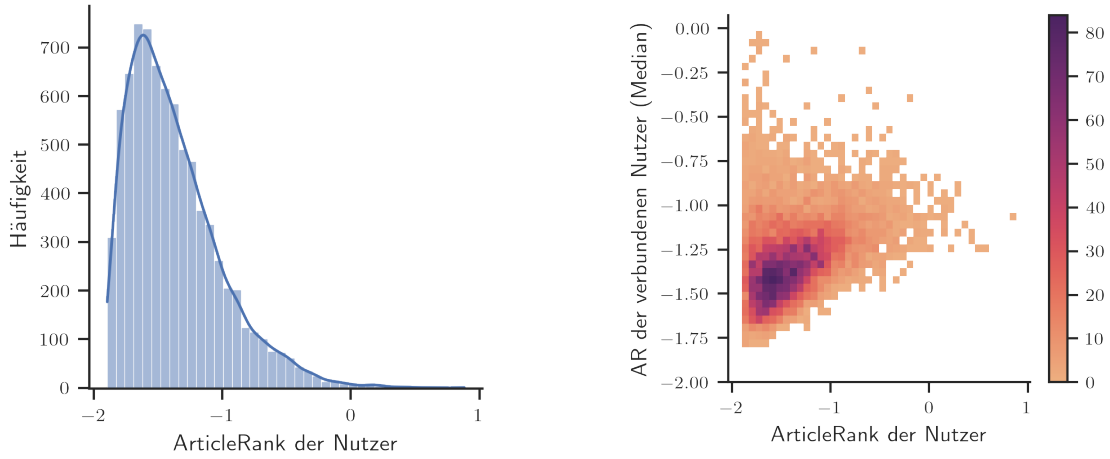


Abbildung 7.12: Verteilung von Nutzern und verbundenen ähnlichen Nutzern. Die meisten Nutzer besitzen einen sehr geringen ArticleRank und sind mit anderen sehr wenig einflussreichen Nutzern verbunden.

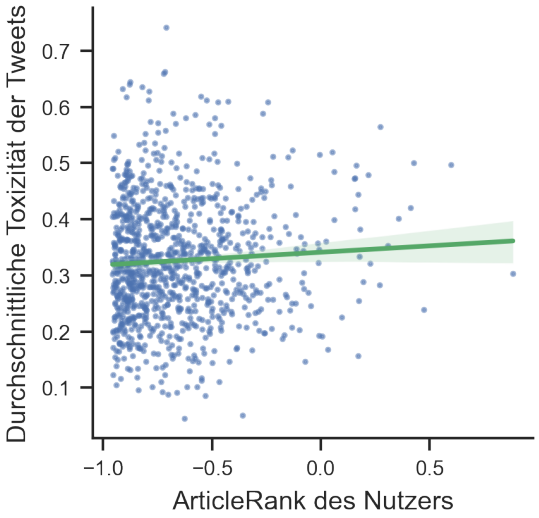


Abbildung 7.13: Bei den 1000 einflussreichsten Nutzern beträgt der Korrelationskoeffizient $r = 0.06$ und die Signifikanz $p = 0.07$

7.3 Modell 3: Tweets

Eine Auswertung von Gruppen ist nicht sinnvoll, da jede Konversation einen in sich geschlossenen Teilgraphen darstellt und somit keine Verbindungen über mehrere Konversationen hergestellt werden können. Als Beispiel wurde die Konversation 1390947260129857537 in Abb. 7.14 dargestellt.

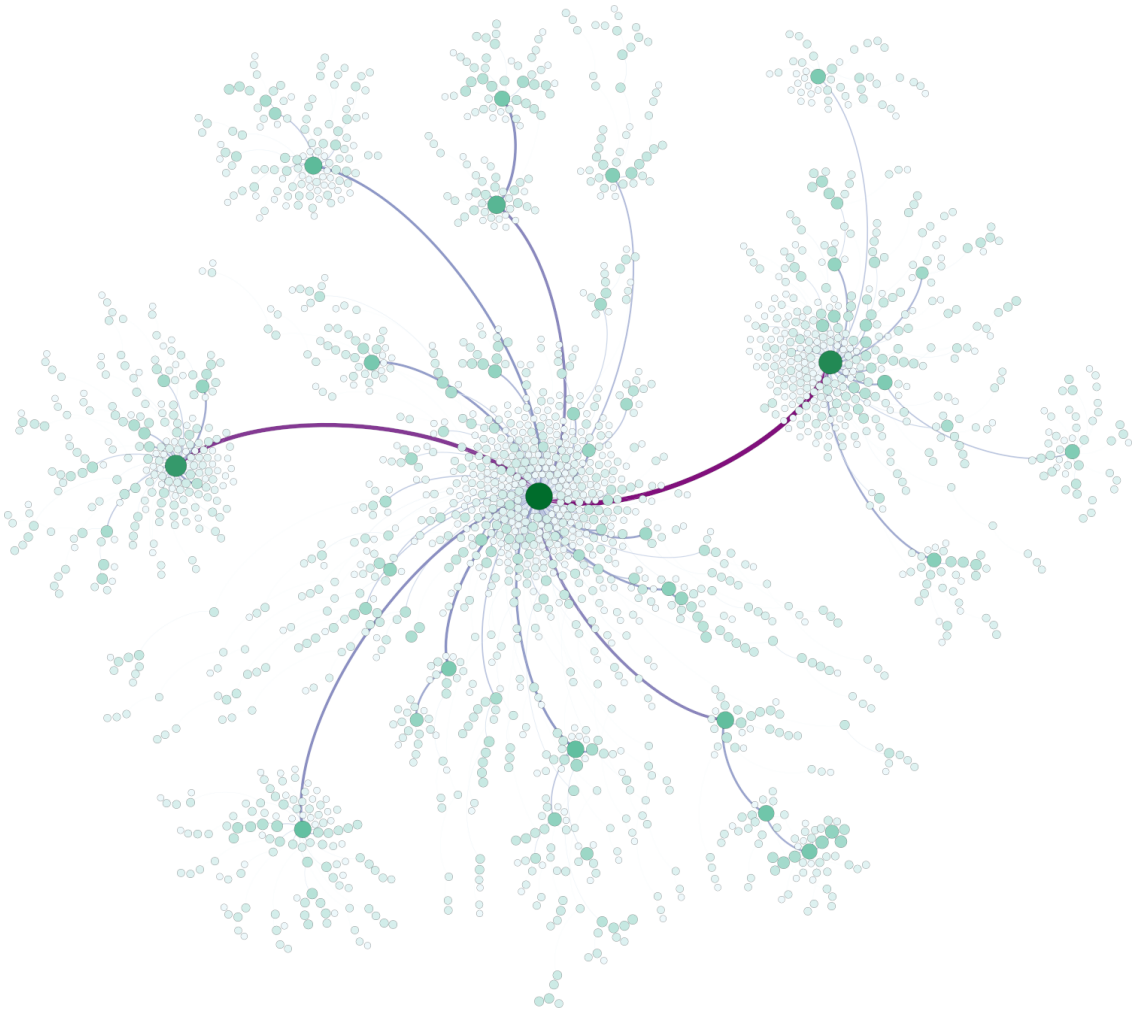


Abbildung 7.14: Abgebildet sind die Tweets der größten Konversation. Es wurden alle Tweets ignoriert, welche 0 *Gefällt mir*-Angaben erhielten oder mehr als fünf Hops vom Main-Tweet entfernt sind. Größe und Sättigung der Knoten stellen den ArtickeRank des Tweets dar. Dicke und Sättigung der Verbindungen stellen den (ausgehenden) Einfluss jedes Tweets dar. Vom Main-Tweet gehen einige große und sehr viele kleine und mittlere Unter-Konversationen ab. Einen hohen ArticleRank erhalten nur Tweets, welche selbst wiederum viele Antworten erhalten.

Dieser Tweet stammt von Annalena Baerbock und lautet:

> Die Äußerung von Boris #Palmer ist rassistisch und abstoßend. Sich nachträglich auf Ironie zu berufen, macht es nicht ungeschehen. Das Ganze reiht sich ein in immer neue Provokationen, die Menschen ausgrenzen und

verletzen. 1/2

Nutzer, deren Tweets im Durchschnitt den höchsten ArticleRank erreichen, sind also besonders in einzelnen Konversationen einflussreich, da sie sehr viele (indirekte) Antworten erhalten. Die Top-Nutzer (mit mindestens 10 Tweets) sind:

1. WELT
2. Janosch Dahmen
3. Philip Plickert
4. AfD_Muenster
5. Ralf Schuler
6. Beatrix von Storch (AfD)
7. Ali Utlu (FDP)
8. Manaf Hassan
9. Benedikt Brechtken
10. Ruprecht Polenz (CDU)

Diese Modellierung scheint besonders Personen des öffentlichen Lebens und Privatpersonen zu erkennen (Abb. 7.15). Die Verteilung des ArticleRank zweigt starke Auffälligkeiten (Abb. 7.16)

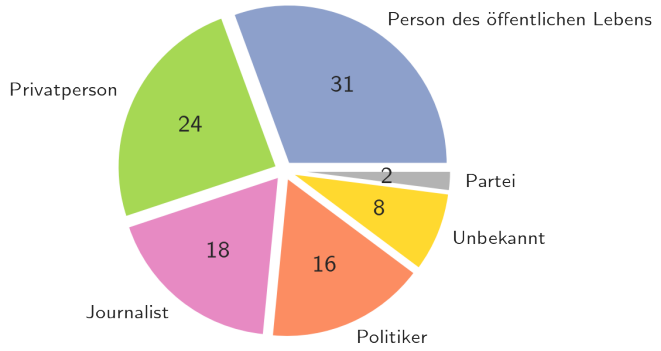


Abbildung 7.15: Kategorien der 50 Top-Nutzer im Tweet-Modell

Auch hier kann eine sehr geringe negative Korrelation zwischen Toxizität und Einfluss fesgestellt werden (Abb. 7.17).

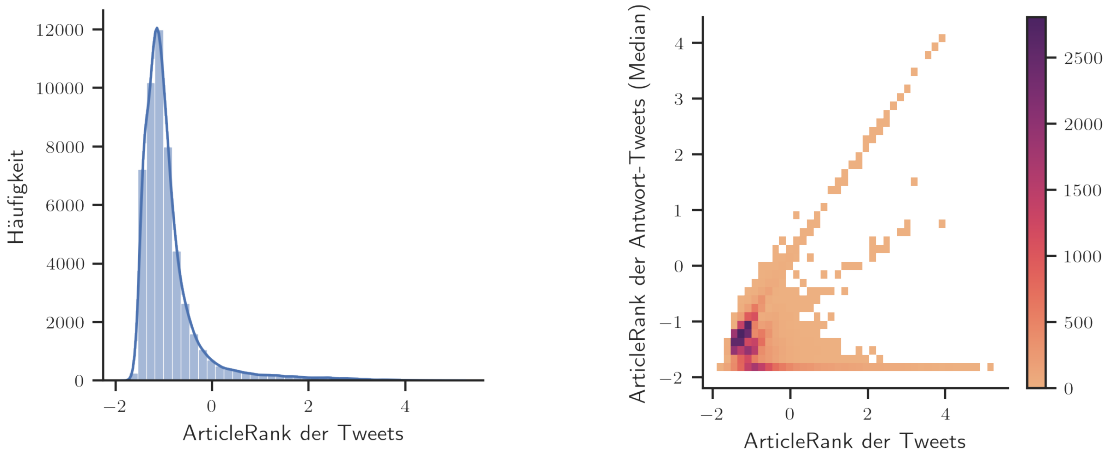


Abbildung 7.16: Der größte Teil der Tweets hat einen sehr geringen ArticleRank und erhielt nur Antworten mit einem sehr geringen ArticleRank. Tweets mit einem hohen bis sehr hohen ArticleRank bekamen entweder eine einzige Antwort mit sehr hohem ArticleRank, mehrere Antworten mit mittlerem ArticleRank oder sehr viele Antworten mit geringem ArticleRank.

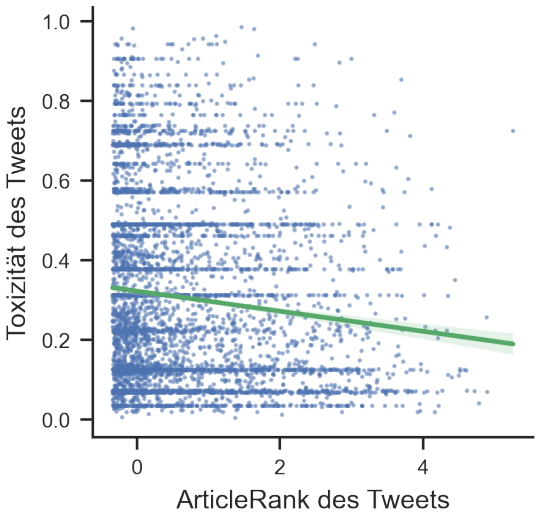


Abbildung 7.17: Bei den 5000 einflussreichsten Tweets beträgt der Korrelationskoeffizient $r = -0,11$ und die Signifikanz $p = 2,44 * 10^{-15}$

8 Fazit und Ausblick

8.1 Fazit

Der Datensatz konnte erfolgreich als Graph in Neo4j modelliert und analysiert werdne. Dieses System ist sehr gut für diesen Zweck geeignet. Es können leicht neue Informationen hinzugefügt und die Graphenstruktur verändert werden. Der zur Verfügung stehende Datensatz ist von hoher Qualität und erlaubt durch die vielen verschiedenen Informationen eine Vielzahl möglicher Analysen.

Aufgrund der Komplexität der Fragestellung wurden drei verschiedene Modelle entwickelt und verglichen. Diese analysierten Antwort-Beziehungen zwischen Nutzern, Nutzer, welche an Konversationen teilnahmen und zuletzt Antwort-Beziehungen zwischen einzelnen Tweets. Das Antwort-Modell erwies sich als am besten geeignet. Es erlaubt die gleichzeitige Analyse von verschiedenen Aspekten des Netzwerkes. Der ArticleRank beschreibt Meinungsführerschaft gut, allerdings sind die durch den Louvain-Algorithmus erkannten Gruppen nicht sehr aussagekräftig. Das Konversationen-Modell kann keine Meinungsführer ausmachen und die Gruppen sind nicht unmittelbar aussagekräftig. Das Tweet-Modell deckt Meinungsführer auf, aber erlaubt keine Auswertung von Gruppen.

Zentralität ist prinzipiell ein passendes Maß, um Meinungsführerschaft zu beschreiben. Dazu wurde eine Auswahl von Algorithmen verglichen. Der numerische Wert des ArticleRank stellt den Grad der Meinungsführerschaft sinnvoll dar. Dabei waren die Unterschiede der Algorithmen sehr ähnlich. Sowohl zwischen PageRank und ArticleRank als auch zwischen dem Wichtungsschema *Anzahl Antworten* und *Einfluss*.

Zur Definition und Beschreibung von Meinungsführern wurde auf medienwissenschaftliche Modelle zurückgegriffen. Die durch den ArticleRank erkannten Meinungsführer entsprechen exakt dieser Definition. Dies bedeutet, dass auf Twitter überwiegend Politiker und politische Aktivisten einflussreich sind. Zusätzlich sind viele Meinungsführer Journalisten, welche über politische Themen berichten. Politische Parteien und Massenmedien werden dagegen fast gar nicht erkannt.

Der ArticleRank alleine beschreibt allerdings nur einen Teil der Informationen des Netzwerkes. Zur Analyse eines Nutzers ist es sinnvoll, dessen Position im Netzwerk genauer zu beschreiben. Dazu wurden die Beziehungen zu anderen Nutzern untersucht. Die Aufteilung der Nutzer in Meinungsführer, Ratsuchende und Inaktive konnte nachgewiesen werden. Circa die Hälfte der Nutzer erhält kaum Antworten und kann daher als Inaktiv gezählt werden. Bei den aktiven Nutzern lassen sich verschiedene Gruppen von Nutzern unterscheiden. Der größte Teil der Nutzer interagiert mit Nutzern mit einem ähnlichen ArticleRank. Die einflussreichsten Nutzer dagegen sind kaum selbst aktiv

und erhalten nur Antworten von deutlich weniger einflussreichen Nutzern. Der Grad der Meinungsführerschaft kann also auch als das Verhältnis zwischen den Nutzern, auf die ein Nutzer antwortet und denen, die ihm antworten.

Es konnte gezeigt werden, dass Gruppen im Sinne der Graphentheorie keine Filterblasen beschreiben. Die erkannten Gruppen umfassen Nutzer, welche ähnliche Themen diskutieren, aber oftmals aus gegensätzlichen Standpunkten aus. Dies liegt daran, dass die Gruppendefinition keine ideologische Ähnlichkeit beschreibt, sondern gemeinsame Aktivität im Netzwerk. Eine weitere Erklärung ist, dass Meinungsführer die Eigenschaft besitzen, in vielen verschiedenen sozialen Gruppen aktiv zu sein. Dies wurde durch die Gruppenerkennung bestätigt.

Durch die Betrachtung des Netzwerkes und der Metadaten lassen sich nur geringe Rückschlüsse auf die Toxizität von Nutzern ziehen. Es wurde gezeigt, dass die stärksten Meinungsmacher weniger toxisch sind als andere Nutzer. Allerdings kann keine Kausalität abgeleitet werden. Weitere Vergleiche zwischen Toxizität und Meinungsführerschaft zeigten keine signifikanten Zusammenhänge.

Zuletzt können durch Werkzeuge wie Gephi kleine bis mittlere Graphen (einige tausend Knoten) effektiv dargestellt werden. Die erlaubt eine intuitive Interpretation der Ergebnisse.

8.2 Ausblick

Diese Arbeit beschränkte sich allein auf die Auswertung der topologischen Daten sowie der Metadaten der Tweets. Eine inhaltliche Auswertung (außer der Bewertung der Toxizität) erfolgte nur stichprobenarttig und manuell. Ein interessanter Ansatz für weitere Forschung ist es, Tweets inhaltlich zu betrachten. So könnte Topic Clustering verwendet werden, um gemeinsame Themen von verschiedenen Nutzern zu erkennen oder Nutzer nach Themen anstatt gemeinsamer Aktivität zu gruppieren. Dadurch kann auch bestimmt werden, welche Themen besonders toxisch diskutiert werden.

Eine weitere Analysemöglichkeit, welche in dieser Arbeit ignoriert wurde, ist die Zeitreihenanalyse. Der Zeitpunkt, wann ein Tweet geschrieben wurde, steht zur Verfügung. Daher wäre es möglich, Themen, Gruppen und Meinungsführer und deren Veränderung im Lauf der Zeit zu bestimmen.

Zur Bestimmung der Toxizität werden ebenfalls andere Methoden benötigt. Die Bewertung von Jigsaw eignet sich zur vergleichbaren Bewertung großer Datenmengen. Allerdings sind weder die Trainingsdaten noch der Algorithmus öffentlich einsehbar. Das ist insbesondere im Hinblick auf die Nachvollziehbarkeit einer forensischen Analyse sehr kritisch. Ebenso wurde die Definition von Toxizität von Jigsaw übernommen, obwohl

diese nicht zwangsläufig der gewünschten Definiton entsprechen muss.

Die Analyse der Meinungsführerschaft sollte sich auch nicht allein auf das Ergebnis eines einzelnen Algorithmus beschränken. Die Zentralität sollte nur benutzt werden, um ein einzelnes Feature zu generieren. Diesen Wert zusammen mit anderen Features (Gruppenkennung, Beziehung zu anderen Knoten, Text-Embedding, etc.) mit einer Standard-Machine-Learning-Pipeline zu untersuchen bietet wahrscheinlich das beste Potential zukünftige Forschungen.

Literatur

AGARWAL, Nitin; DOKOOHAKI, Nima; TOKDEMIR, Serpil (Hrsg.), 2019. *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer International Publishing. Abger. unter DOI: 10.1007/978-3-319-94105-9 (siehe S. 26).

ALLEN, David u. a., 2019. Understanding Trolls with Efficient Analytics of Large Graphs in Neo4j. In: T. GRUST ET AL. (Hrsg.). *Datenbanksysteme für Business, Technologie und Web*. Gesellschaft für Informatik, Bonn, S. 377–396. Abger. unter DOI: 10.18420/BTW2019-23 (siehe S. 47, 48).

BASTIAN, Mathieu; HEYMANN, Sebastien; JACOMY, Mathieu, 2009. Gephi : An Open Source Software for Exploring and Manipulating Networks. In: *International AAAI Conference on Weblogs and Social Media* (siehe S. 6, 31).

BHAT, Sajid Yousuf; ABULAISH, Muhammad, 2017. A Unified Framework for Community Structure Analysis in Dynamic Social Networks. In: *Hybrid Intelligence for Social Networks*. Springer International Publishing, S. 77–97. Abger. unter DOI: 10.1007/978-3-319-65139-2_4 (siehe S. 5, 12).

BRANDT, Mathias, 2019. *Onlinehass ist vor allem rechtsradikal* [online]. Hrsg. von STATISTA. 2019-11-26 [besucht am 03. 10. 2021]. Abger. unter: https://de.statista.com/infografik/20115/polizeilich-erfasste-hasskommentare-in-deutschland/ (siehe S. 3).

BRIN, Sergey; PAGE, Lawrence, 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. Jg. 30, Nr. 1-7, S. 107–117. Abger. unter DOI: 10.1016/s0169-7552(98)00110-x (siehe S. 21).

BRONSTEIN, Michael, 2020. *Deep learning on graphs: successes, challenges, and next steps | Graph Neural Networks* [online]. Hrsg. von INTELLIGENT SYSTEMS CONFERENCE (INTELLISYS). 2020-09-04 [besucht am 07. 08. 2021]. Abger. unter: https://www.youtube.com/watch?v=PLGcx65MhCc (siehe S. 26).

BSI, 2021. *IT-Grundschutz-Kompendium*. Hrsg. von BUNDESAMT FÜR SICHERHEIT IN DER INFORMATIONSTECHNIK. Köln: Reguvis Fachmedien GmbH. ISBN 978-3-8462-0906-6 (siehe S. 11).

BSI, 2011. *Leitfaden „IT-Forensik“*. Bundesamt für Sicherheit in der Informationstechnik (siehe S. 11, 12).

CASWELL, Thomas A u. a., 2021. *matplotlib/matplotlib: REL: v3.4.2*. Zenodo. Abger. unter DOI: 10.5281/ZENODO.4743323 (siehe S. 6).

DRESSLER, Matthias; TELLE, Gina, 2009. *Meinungsführer in der interdisziplinären Forschung*. Gabler, Betriebswirt.-Vlg. ISBN 3834914762. Verfügbar unter: https://www.ebook.de/de/product/7973921/matthias_dressler_gina_telle_meinungsfuehrer_in_der_interdisziplinaeren_forschung.html (siehe S. 3, 5, 9, 10).

DU, Siying; GREGORY, Steve, 2016. The Echo Chamber Effect in Twitter: does community polarization increase? In: *Studies in Computational Intelligence*. Springer International Publishing, S. 373–378. Abger. unter DOI: 10.1007/978-3-319-50901-3_30 (siehe S. 8).

DUDENREDAKTION (Hrsg.), [o. D.]. *Duden online* [online] [besucht am 17. 07. 2021]. Abger. unter: https://www.duden.de (siehe S. 7, 19, 23, 25).

FACEBOOK, [o. D.]. *Community Standards Enforcement Report: Hate Speech* [online] [besucht am 30. 09. 2021]. Abger. unter: https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/ (siehe S. 3).

FERNQUIST, Johan u. a., 2020. Twitter Bots and the Swedish Election. In: Springer International Publishing, S. 141–163. Abger. unter DOI: 10.1007/978-3-030-41251-7_6 (siehe S. 3).

GEISE, Stephanie, 2020. Theorieansätze und Hypothesen in der Medienpädagogik: Meinungsführer und der „Flow of Communication“. In: *Handbuch Medienpädagogik*. Springer Fachmedien Wiesbaden, S. 1–8. Abger. unter DOI: 10.1007/978-3-658-25090-4_38-1 (siehe S. 9).

GROVER, Aditya; LESKOVEC, Jure, 2016. node2vec: Scalable Feature Learning for Networks. Abger. unter arXiv: 1607.00653 [cs.SI] (siehe S. 26).

GUARINO, Stefano u. a., 2020. Characterizing networks of propaganda on twitter: a case study. *Applied Network Science*. Jg. 5, Nr. 1. Abger. unter DOI: 10.1007/s41109-020-00286-y (siehe S. 12, 48).

HAGBERG, Aric A.; SCHULT, Daniel A.; SWART, Pieter J., 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In: *Proceedings of the 7th Python in Science Conference*, S. 11–15 (siehe S. 31).

HODLER, Amy; NEEDHAM, Mark, 2019. *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly Media. ISBN 9781492057819 (siehe S. 5, 13, 19, 24, 31).

HODLER, Amy; NEEDHAM, Mark, 2021. *Graph Data Science (GDS): for dummies*. John Wiley & Sons, Inc. ISBN 9781119746041 (siehe S. 5, 19, 31).

HUNTER, John D., 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*. Jg. 9, Nr. 3, S. 90–95. Abger. unter DOI: 10.1109/MCSE.2007.55.

INHOFFEN, Lisa, 2019. *Hassrede im Netz ist vor allem für AfD-Wähler kein No-Go*. Hrsg. von YOUGOV, STATISTA. 2019-10-21. Verfügbar unter: https://yougov.de/news/2019/10/21/hassrede-im-netz-ist-vor-allem-fur-afd-wahler-kein/ (siehe S. 3).

JIGSAW; GOOGLE (Hrsg.), [o. D.]. *Perspective*. Verfügbar unter: https://www.perspectiveapi.com/ (siehe S. 40).

KAISER, Carolin, 2012. *Business Intelligence 2.0*. Gabler Verlag. Abger. unter DOI: 10.1007/978-3-8349-3990-6 (siehe S. 5).

KATZ, Elihu, 1957. The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opinion Quarterly*. Jg. 21, Nr. 1, Anniversary Issue Devoted to Twenty Years of Public Opinion Research, S. 61. Abger. unter DOI: 10.1086/266687 (siehe S. 21, 27).

KENT, Karen u. a., 2006. Guide to Integrating Forensic Techniques into Incident Response: Recommendations of the National Institute of Standards and Technology. In: *Reports on Computer Systems Technology*. Bd. Special Publication 800-86 (siehe S. 12).

AL-KHATEEB, Samer; AGARWAL, Nitin, 2019. *Deviance in Social Media and Social Cyber Forensics*. Springer International Publishing. Abger. unter DOI: 10.1007/978-3-030-13690-1 (siehe S. 4, 11, 12).

KITAJIMA, Yuzuki; OTAKE, Kohei; NAMATAME, Takashi, 2021. Analysis of User Relationships on Cooking Recipe Site Using Network Structure. In: *Social Computing and Social Media: Experience Design and Social Network Analysis*. Springer International Publishing, S. 284–300. Abger. unter DOI: 10.1007/978-3-030-77626-8_19 (siehe S. 47).

LABUDDE, Dirk; CZERNER, Frank; SPRANGER, Michael, 2017. Einführung. In: *Forensik in der digitalen Welt*. Springer Berlin Heidelberg, S. 1–23. Abger. unter DOI: 10.1007/978-3-662-53801-2_1 (siehe S. 10).

LESKOVEC, Jure; SOSIČ, Rok, 2016. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Transactions on Intelligent Systems and Technology*. Jg. 8, Nr. 1, S. 1–20. Abger. unter DOI: 10.1145/2898361 (siehe S. 31).

LEX, Alexander u. a., 2014. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*. Jg. 20, Nr. 12, S. 1983–1992. Abger. unter DOI: 10.1109/tvcg.2014.2346248.

LI, Jiang; WILLETT, Peter, 2009. ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks. *Aslib Proceedings*. Jg. 61, Nr. 6, S. 605–618. Abger. unter DOI: 10.1108/00012530911005544 (siehe S. 22).

LU, Hao; HALAPPANAVAR, Mahantesh; KALYANARAMAN, Ananth, 2014. Parallel Heuristics for Scalable Community Detection. Abger. unter arXiv: 1410.1237 [cs.SI] (siehe S. 24).

LU, Linyuan u. a., 2011. Leaders in Social Networks, the Delicious Case. *PLoS ONE 6(6): e21202 (2011)*. Abger. unter DOI: 10.1371/journal.pone.0021202 (siehe S. 22).

LUBARSKI, Paweł; MORZY, Mikołaj, 2014. @Rank: Personalized Centrality Measure for Email Communication Networks. In: *Lecture Notes in Social Networks*. Springer International Publishing, S. 209–225. Abger. unter DOI: 10.1007/978-3-319-05912-9_10 (siehe S. 22).

LYU, Liang u. a., 2021. Centrality with Diversity. Abger. unter DOI: 10.1145/3437963.3441789 (siehe S. 22).

MCINNES, Leland; HEALY, John; MELVILLE, James, 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Abger. unter arXiv: 1802.03426 [stat.ML] (siehe S. 79).

MCKINNEY, Wes, 2010. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. SciPy. Abger. unter DOI: 10.25080/majora-92bf1922-00a.

MERKL SASAKI, Bryce; CHAO, Joy; HOWARD, Rachel, 2018. *Graph Databases for Beginners*. Neo4j, Inc (siehe S. 5).

MRSIC, Leo; ZAJEC, Srecko; KOPAL, Robert, 2019. Appliance of Social Network Analysis and Data Visualization Techniques in Analysis of Information Propagation. In: *Intelligent Information and Database Systems*. Springer International Publishing, S. 131–143. Abger. unter DOI: 10.1007/978-3-030-14802-7_11 (siehe S. 5, 12).

NEO4J, INC, OPENCYPHER (Hrsg.), 2021. Cypher Query Language Reference,Version 9 (siehe S. 32).

NEO4J, INC. (Hrsg.), 2021. *Neo4j: Graphs for Everyone* [online]. Version 4.2.6 [besucht am 21. 05. 2021]. Abger. unter: https://neo4j.com (siehe S. 31).

NICHOLAS, Tom, 2021. *Veritasium: A Story of YouTube Propaganda* [online]. 2021-10-20 [besucht am 26. 10. 2021]. Abger. unter: https://www.youtube.com/watch?v=CM0aohBfUTc&t=2s (siehe S. 4).

PEDREGOSA, F. u. a., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. Jg. 12, S. 2825–2830 (siehe S. 6).

POTTHOFF, Matthias (Hrsg.), 2016. *Schlüsselwerke der Medienwirkungsforschung*. Springer Fachmedien Wiesbaden. Abger. unter DOI: 10.1007/978-3-658-09923-7 (siehe S. 5).

RAJEH, Stephany u. a., 2021. Characterizing the Interactions Between Classical and Community-aware Centrality Measures in Complex Networks. *Sci Rep 11, 10088 (2021)*. Abger. unter DOI: 10.1038/s41598-021-89549-x (siehe S. 22).

REBACK, Jeff u. a., 2021. *pandas-dev/pandas: Pandas 1.3.0*. Zenodo. Abger. unter DOI: 10.5281/ZENODO.3509134 (siehe S. 6).

ROBINSON, Ian; WEBBER, Jim; EIFREM, Emil, 2015. *Graph Databases: New opportunities for connected data*. 2. Aufl. Sebastopol, CA: O'Reilly. ISBN 9781491932001 (siehe S. 5).

SAXENA, Akrati; IYENGAR, Sudarshan, 2020. Centrality Measures in Complex Networks: A Survey. Abger. unter arXiv: 2011.07190 [cs.SI] (siehe S. 20).

SCHACH, Annika; LOMMATZSCH, Timo (Hrsg.), 2018. *Influencer Relations*. Springer Fachmedien Wiesbaden. Abger. unter DOI: 10.1007/978-3-658-21188-2 (siehe S. 4, 5, 10).

SHAFIQ, M. Z. u. a., 2013. Identifying Leaders and Followers in Online Social Networks. Jg. 31, Nr. 9, S. 618–628. Abger. unter DOI: 10.1109/jsac.2013.sup.0513054 (siehe S. 10).

SHIN, Soo-jin u. a., 2014. Study on Relation between Social Circles and Communities in Facebook Ego Networks. In: *Lecture Notes in Electrical Engineering*. Springer Berlin Heidelberg, S. 567–572. Abger. unter DOI: 10.1007/978-3-642-41671-2_72 (siehe S. 24).

SKIENA, Steven S., 2020. *The Algorithm Design Manual*. Springer International Publishing. Abger. unter DOI: 10.1007/978-3-030-54256-6 (siehe S. 47).

SPRANGER, Michael u. a., 2020. Measuring Competence: Improvements to Determine the Degree of Opinion Leadership in Social Networks. *International Journal on Advances in Internet Technology*. Jg. 13, Nr. 3–4, S. 97–109 (siehe S. 22, 50, 56).

STATISTA, STATISTA DIGITAL MARKET OUTLOOK (Hrsg.), 2019. *Digital Market Outlook* [online]. 2019-02 [besucht am 03. 10. 2021]. Abger. unter: https://de.statista.com/statistik/daten/studie/554909/umfrage/anzahl-der-nutzer-sozialer-netzwerke-in-deutschland/ (siehe S. 3).

TADDICKEN, Monika, 2015. The People's Choice. How the Voter Makes Up His Mind in a Presidential Campaign. In: Springer Fachmedien Wiesbaden, S. 25–36. Abger. unter DOI: 10.1007/978-3-658-09923-7_3 (siehe S. 4).

THE IGRAPH CORE TEAM, 2021. *igraph: The network analysis package* [online]. Version 0.9.1 [besucht am 30. 04. 2021]. Abger. unter: https://igraph.org/ (siehe S. 31).

TWITTER, 2019. *Q3 2019 Letter to Shareholders*. 2019-10-24. Techn. Ber. (siehe S. 3).

TWITTER, [o. D.]. *Rules Enforcement* [online] [besucht am 30. 09. 2021]. Abger. unter: https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec (siehe S. 3).

TWITTER, 2021. Twitter Netzwerkdurchsetzungsbericht: Januar - Juni 2021 (siehe S. 3).

VAN ROSSUM, Guido, 2019. *Python Patterns - Implementing Graphs* [online]. 2019-11-08 [besucht am 01. 05. 2021]. Abger. unter: https://www.python.org/doc/essays/graphs/ (siehe S. 30).

VIRTANEN, Pauli u. a., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. Jg. 17, Nr. 3, S. 261–272. Abger. unter DOI: 10.1038/s41592-019-0686-2 (siehe S. 6).

WAN, Zelin u. a., 2020. A Survey on Centrality Metrics and Their Implications in Network Resilience. Abger. unter arXiv: 2011.14575 [cs.SI] (siehe S. 20).

WASKOM, Michael, 2021. seaborn: statistical data visualization. *Journal of Open Source Software*. Jg. 6, Nr. 60, S. 3021. Abger. unter DOI: 10.21105/joss.03021 (siehe S. 6).

WASSERMAN, Stanley; FAUST, Katherine, 1995. *Social Network Analysis 1ed*. Cambridge University Press. ISBN 0521387078. Verfügbar unter: https://www.ebook.de/de/product/3719721/stanley_wasserman_katherine_faust_social_network_analysis_1ed.html (siehe S. 15).

WEBBER, Jim; BRUGGEN, Rik Van; EIFREM, Emil, 2020. *Graph Databases*. John Wiley & Sons, Inc. ISBN 978-1-119-74579-2 (siehe S. 13).

WENG, Jianshu u. a., 2010. TwitterRank. In: *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*. ACM Press. Abger. unter DOI: 10.1145/1718487.1718520 (siehe S. 22).

XU, Shuqi u. a., 2020. Unbiased evaluation of ranking metrics reveals consistent performance in science and technology citation data. Abger. unter DOI: 10.1016/j.joi.2019.101005 (siehe S. 21).

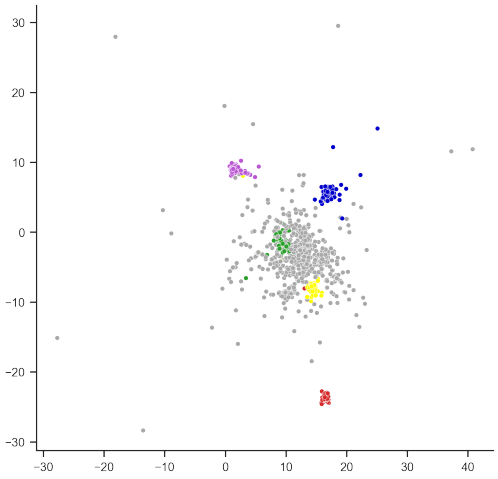
ZHANG, Ziwei; CUI, Peng; ZHU, Wenwu, 2020. Deep Learning on Graphs: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, S. 1–1. Abger. unter DOI: 10.1109/tkde.2020.2981333 (siehe S. 5, 26).

ZWEIG, Katharina A., 2016. *Network Analysis Literacy: a practical approach to the analysis of networks*. Springer-Verlag KG. ISBN 3709107415. Verfügbar unter: https://www.ebook.de/de/product/27812630/katharina_a_zweig_network_analysis_literacy.html (siehe S. 5, 15, 20, 21, 25, 26, 47).

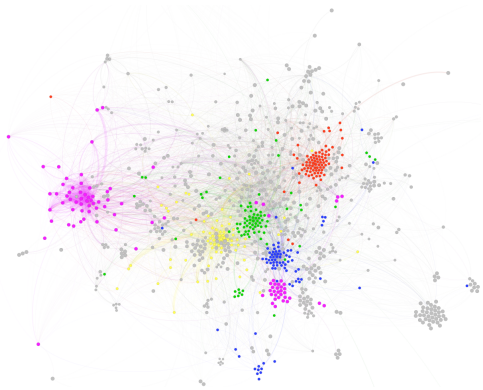
ZWEIG, Katharina A.; DEUSSEN, Oliver; KRAFFT, Tobias D., 2017. Algorithmen und Meinungsbildung. *Informatik-Spektrum*. Jg. 40, Nr. 4, S. 318–326. Abger. unter DOI: 10.1007/s00287-017-1050-5 (siehe S. 7).

Anhang A: Node Embeddings

Um die Leistungsfähigkeit und Anwendbarkeit von maschinellen Lernverfahren zu demonstrieren wurden mit dem Node2vec-Algorithmus Embeddings der Graph-Modelle erstellt. Dieses Embedding wurde anschließend zur grafischen Darstellung mit *Uniform Manifold Approximation and Projection* (UMAP) (McInnes u. a. 2018) auf zwei Dimensionen abgebildet (Abb. A.1 und A.2). Node2vec verwendet Random Walks zweiter Ordnung. Das bedeutet, dass in jedem Schritt der zuletzt besuchte Knoten bekannt ist. Für jeden Knoten werden mit einem Random Walk die benachbarten Knoten erkundet, um nahe beieinander stehende Knoten zu finden. Mit zwei Parametern (p, q) kann dabei zwischen einer Breitensuche und einer Tiefensuche abgewogen werden. In Abhängigkeit von p und q werden unterschiedliche Informationen über die Nachbarschaft der Knoten abgebildet. Auch bei denn Tweets kann durch eine solche Projektion eine Struktur dargestellt werden (Abb. A.3).



(a) Darstellung des Knoten-Embeddings



(b) Darstellung des Graphen mit Gephi

Abbildung A.1: Vergleich der Abbildung des Graphen mit der Abbildung des Embeddings. Für die linke Grafik wurde zuerst mit dem Node2vec-Algorithmus ein 128-dimensionales Embedding der Nutzer erstellt. Auch hier stellt die Farbe die Gruppenzugehörigkeit dar. Die Gruppenstruktur ist auch im Embedding-Raum erhalten geblieben.

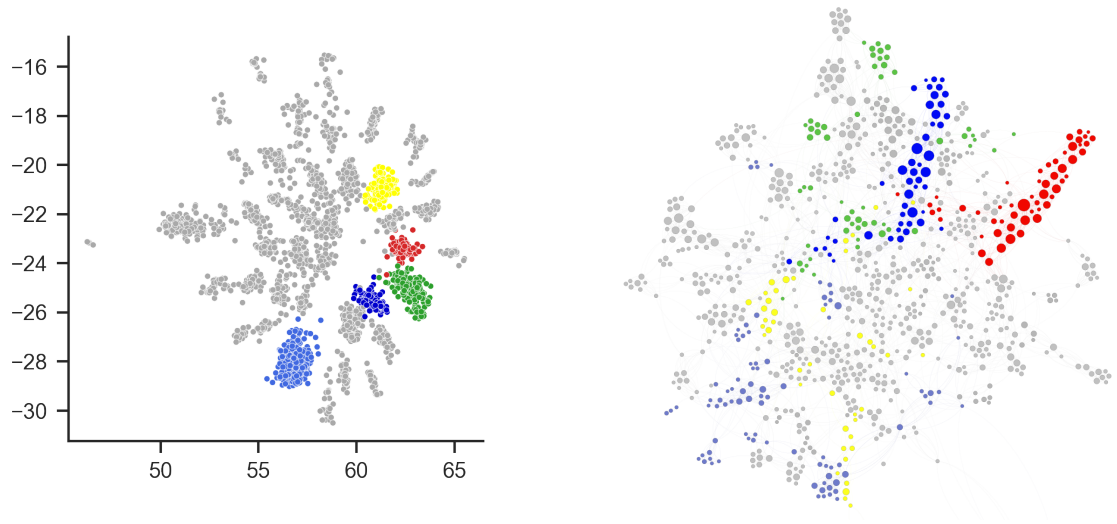


Abbildung A.2: Auch beim Konversationen-Modell konnte das Node Embedding die Struktur der Gruppen gut erfassen.

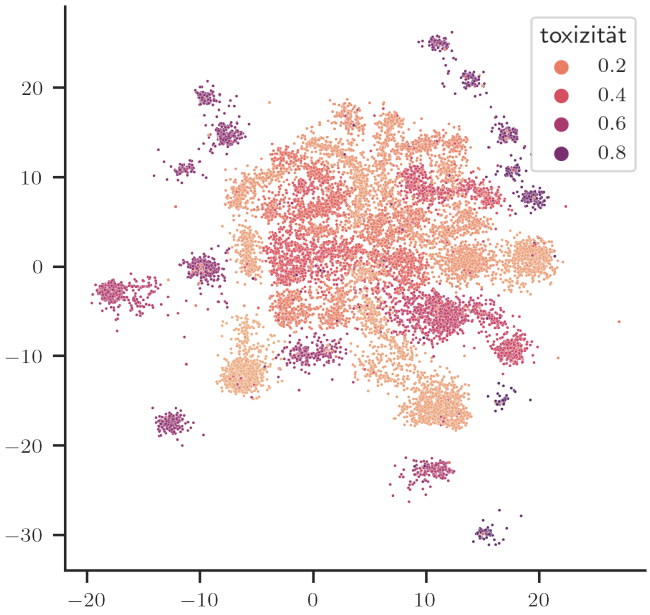


Abbildung A.3: Zweidimensionale Darstellung der Tweets. Für jeden Tweet wurden sieben Dimensionen betrachtet: „Gefällt mir"-Angaben, Antworten, Zitat-Tweets, Retweets, ArticleRank des Autors und ArticleRank des Tweets. Ob die größern Cluster jeweils einen hohen Ausschlag im Histogram der Toxizitätswerte darstellen wurde nicht überprüft.

Anhang B: Programmcode-Beispiele

```python
from neo4j import GraphDatabase
import pandas as pd

bolt_uri = "bolt://localhost:7687"
auth = ("neo4j", "passwort")
driver = GraphDatabase.driver(uri=bolt_uri, auth=auth)

query = """
    MATCH
      (t:Tweet)
    WHERE
      t.likeCount > 100
    RETURN
      t as Tweet
"""

with driver.session(database="neo4j") as session:
    result = session.run(query)
    result = pd.DataFrame([dict(record) for record in result])
    print (result)
```

Listing B.1: Ausführung einer Cypheranfrage mittels Python und dem offiziellen Python-Treiber von Neo4j. Abgefragt werden alle Tweets mit mehr als 100 „Gefällt mir"-Angaben. Das Ergebnis wird in ein Pandas-DataFrame gespeichert.

```
CALL gds.articleRank.write({
    nodeProjection: "Nutzer",
    relationshipProjection: "ANTWORTETE_AUF",
    relationshipProperties: "anzahlAntworten",
    relationshipWeightProperty: "anzahlAntworten",
    writeProperty: "articleRankAntworten"
    })
```

Listing B.2: Berechnung des ArticleRanks der Nutzer. Es wird nur der Knotentyp *Nutzer* und der Kantentyp *ANTWORTETE_AUF* betrachtet. Die Eigenschaft *anzahlAntworten* dient als Gewicht der Beziehung.

Anhang C: Tabellen

Text	Vorkommen
@Karl_Lauterbach Gute Besserung!	797
@Karl_Lauterbach Gute Besserung	538
@SHomburg #RKIGate	263
@LutzvanderHorst Gute Besserung!	222
@MickyBeisenherz Gute Besserung!	121

Tabelle C.1: Text und Häufigkeit der fünf häufigsten Tweet-Texte. Karl Lauterbach, Micky Beisenherz sowie Lutz van der Horst waren an Corona erkrankt und bekamen deshalb Genesungswünsche. #RKIgate bezog sich auf angeblich gefälschte Zahlen von Covid19-Infizierten vom Robert-Koch-Institut.

Nutzer		Tweets
Twitterhandle	Anzeigename	
plus_eins_plus	Joe	3.328
EmmaWag68768896	Emma Wagner	2.305
Dunkelfluegel	EddieErpel	1.564
RenaadeS	R. S. Klarname	1.476
ChakerRonai	Ronai Chaker	1.180

Tabelle C.2: Die fünf Nutzer mit den meisten Tweets.

Nutzer		Erhaltende Antworten (Anzahl)
Twitterhandle	Anzeigename	
Karl_Lauterbach	Karl Lauterbach	46.827
janboehm	Jan Böhmermann	25.596
reitschuster	Boris Reitschuster	20.181
ABaerbock	Annalena Baerbock	17.922
Markus_Soeder	Markus Söder	15.271

Tabelle C.3: Die fünf Nutzer mit den meisten erhaltenen Antworten.

Nutzer		Verhältnis Antworten/Tweet
Twitterhandle	Anzeigename	
LutzvanderHorst	Lutz van der Horst	864
ABaerbock	Annalena Baerbock	663
jensspahn	Jens Spahn	530
Afelia	Marina Weisband	426
Markus_Soeder	Markus Söder	424

Tabelle C.4: Nutzer mit dem höchsten Verhältnis aus erhaltenen Antworten und geschriebenen Tweets (gerundet)

Tweet	„Gefällt mir"-Angaben
Lutz van der Horst	
#allesdichtmachen - Ich bin auch dabei https://t.co/AytVnvElmh	50.372
Karl Lauterbach	
Liebe Frau @Alice_Weidel. Jeder Platz auf der SPD Liste ist mehr Wert als Spitzenplätze auf der Liste einer Partei, die Hass und Hetze in den Bundestag getragen hat. https://t.co/9eudYz7Iwp	35.547
Karl Lauterbach	
Ich bringe nur das Nötigste zur Anzeige. Drohungen, Aufrufe zur Gewalt und Straftaten, menschenverachtende Beleidigungen der letzten Wochen. Trotzdem habe ich gerade wieder 59 Anzeigen unterschrieben. Ich weiss, dass es auch Wissenschaftlern so geht, die gegen Pandemie kämpfen https://t.co/RbfDj9smOj	32.936
Karl Lauterbach	
(1) In eigener Sache: ich habe in den letzten Tagen nicht getwittert und auch Medientermine abgesagt. Bevor spekuliert wird: Musste mich kurzfristig einer Augen Operation unterziehen. Hoffe auf baldige Genesung. Werde 1 Woche kürzer treten. Trotzdem verfolge ich alles	27.485
Annalena Baerbock	
Die Äußerung von Boris #Palmer ist rassistisch und abstoßend. Sich nachträglich auf Ironie zu berufen, macht es nicht ungeschehen. Das Ganze reiht sich ein in immer neue Provokationen, die Menschen ausgrenzen und verletzen. 1/2	23.817

Tabelle C.5: Die fünf Tweets mit den meisten „Gefällt mir"-Angaben.

Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 30. November 2021