DHBW
Duale Hochschule
Baden-Württemberg
Stuttgart
Campus Horb

Untersuchung, Implementierungen und Bewertung von Graph-Metriken

Studienarbeit

im Studiengang Informatik

an der Dualen Hochschule Baden-Württemberg Stuttgart, Campus Horb am Neckar

von

Benedict Weichselbaum

12. November 2020

Bearbeitungszeitraum	28.09.2020 - 31.05.2021
Matrikelnummer, Kurs	6275457, TINF2018
Betreuer & Gutachter	Prof. Dr. ing. Olaf Herden

Erklärung

Ich versichere hiermit, dass ich meine Studienarbeit mit dem Thema *Graphen: Metriken und Ähnlichkeit* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.
Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Nürnberg, 12. November 2020

Benedict Martin Weichselbaum

Abstract

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

1 Einleitung

1.1 Motivation für die Studienarbeit

Graphen sind einer der wichtigsten Datenstrukturen der Informatik. Warum kann man das sagen? In seinem Buch „Algorithmische Graphentheorie“ nennt Volker Turau, Professor an der Universität Hamburg-Harburg, den Grund dafür:

> *Graphen sind die in der Informatik am häufigsten verwendete Abstraktion. Jedes System, welches aus diskreten Zuständen oder Objekten und Beziehungen zwischen diesen besteht, kann als Graph modelliert werden.* [Tur04]

Diese netzartigen Strukturen können dabei die verschiedensten Konstrukte repräsentieren. Dazu zählen Straßennetze, Computernetzwerke, elektrische Schaltungen aber auch zum Beispiel chemische Moleküle. [Tit19]

Um Graphen zu beschreiben und zu charakterisieren, haben sich über die Zeit zahlreiche Metriken, bzw. Eigenschaften für diese herausgebildet („graph properties“ [Lov12]). Das heißt, einem Graphen können gewisse Kennzahlen zugeordnet werden, die ihn auszeichnen. Auch diese Metriken sind, wie die Graphen selbst, meist praktisch anwendbar. Zum Beispiel in der Untersuchung von Netzwerken [EK13].

Diese Studienarbeit soll nun diese Metriken genauer untersuchen. Hierbei ist es zunächst wichtig die verschiedenste Metriken vorzustellen und zu erläutern. Dabei ist es auch wichtig herauszufinden, wie verbreitet diese Metriken sind und inwieweit die jeweiligen Kennzahlen zu bewerten sind. Des Weiteren soll auf Basis der Metriken auch der Begriff der Ähnlichkeit von Graphen aufgegriffen werden.

Neben einer theoretischen Betrachtung der Graphmetriken soll auch eine Implementierung stattfinden. Es ist dabei das Ziel, mithilfe von Graphdatenbanken die jeweiligen Metriken umzusetzen und diese miteinander zu Vergleichen.

In einem Weiteren Teil ist außerdem noch darauf einzugehen, welche Anwendung die gezeigten Metriken haben, um den praktischen Nutzen der Thematik aufzuzeigen.

1.2 Fragestellungen

Auf Basis dieser Motivation können nun auch die konkreten Fragestellungen formuliert werden, die diese Arbeit betrachten soll. Insgesamt sollen vier wissenschaftliche Fragen beantwortet werden.

1. Welche Graph-Metriken gibt es und wie sind diese zu ermitteln und zu kategorisieren?

Hierzu gehört, wie bereits erwähnt die Vorstellung der einzelnen Metriken, aber auch eine Kategorisierung in Rubriken, um Metriken besser voneinander abzugrenzen, da diverse Metriken höchst unterschiedliche Aussagen über einen Graphen treffen. Es wird auch darauf eingegangen welche Motivation hinter den jeweiligen Metriken steht. Bei der Beantwortung dieser Frage soll außerdem auch darauf eingegangen werden, inwieweit die beschriebene Metrik in bestimmten Mathematikbibliotheken wie „Sage Math“ oder „Wolfram“ vorkommen.

2. Wie sind die vorgestellten Metriken zu bewerten?

In diesem Abschnitt soll es vor allem darum gehen, die vorgestellten Metriken dahingehend zu bewerten, wie „schwer“ es ist, sie zu ermitteln. Außerdem soll bei der Bewertung auch auf die Verbreitung eingegangen werden.

3. Was beschreibt der Ähnlichkeitsbegriff bei Graphen?

Basierend auf Graph-Metriken lässt sich auch ermitteln, ob zwei Graphen Ähnlichkeiten aufweißen [WM19]. Auch auf diesen Aspekt soll die Arbeit bezug nehmen und dabei ein Anwendungs-Beispiel konstruieren.

4. Wie können die vorgestellten Metriken in Graphdatenbanken verwendet werden, bzw. implementiert werden?

Auf die theoretische Betrachtung der Graph-Metriken folgt dann ein praktischer Teil, der behandeln soll, wie sich die Metriken in bekannten Graphdatenbanken umsetzen lassen, bzw. umgesetzt wurden. Dabei ist es wichtig herauszufinden welche Graphkennzahlen bereits teil der Graphdatenbank-Lösungen sind, bzw. welche Metriken selbst umgesetzt werden müssen.

5. Wie sind die jeweiligen Implementierungen zwischen und innerhalb der Graphdatenbanken zu bewerten?

Folgend auf die Implementierung, ist es noch wichtig zu verstehen, wie diese Umsetzungen zu betrachten sind. Dabei wird vor allem ein Fokus auf das Thema Performance und Skalierung gelegt.

6. Welche Anwendungen gibt es für Graph-Metriken und den Vergleich von Graphen (Ähnlichkeit)?

Als letztes soll sich die Studienarbeit mit praktischen Beispielen beschäftigen. Es ist dabei wichtig zu verstehen, welchen konkreten Nutzen die gezeigten Kennzahlen für Graphen in modernen Anwendungsszenarien haben.

2 Graph-Metriken

Dieser erste Teil der Arbeit wird sich nun ausführlich mit einer weiten Reihe an Graph-Metriken beschäftigen. Hierbei sollen die ersten zwei Fragestellungen der Arbeit genau beantwortet werden. Zur jeweiligen Vorstellung einer Graph-Metrik sollen dabei die folgengen Punkte erläutert werden:

- Was drückt die Metrik aus (Definition)?
- Welche Motivation hat die Metrik?
- Inwieweit ist die Metrik verbreitet? Zum Beispiel in der Literatur oder in Mathematikbibliotheken.
- Wie ist die Metrik im Bezug auf den Rechenaufwand zu bewerten?

Es ist noch zu erwähnen, dass alle im folgenden vorgestellten Metriken über die einzelnen Sektionen der Arbeit in Kategorien eingeteilt sind.

Darüber hinaus ist noch eine grundsätzliche Notationen während der Arbeit zu klären: Ein **Graph G** ist ein Paar bestehend aus **Knoten V** und **Kanten E**.

$$G = (V, E),\ wobei\ E \subseteq V \times V$$

Für V können wir auch V(G) schreiben, für E auch E(G). [Die00] V ist dabei Englisch und bedeutet „Vertices“, E steht für „Edges“. Wenn es um die Datenstrukturen von Graphen geht, kommen im Rahmen dieser Arbeit hauptsächlich Adjazenzmatrizen und Adjazenzlisten zum Einsatz. Allerdings können bei Bedarf auch Inzidenzen (Beziehung zwischen Knoten und Kanten) im Graphen eine Rolle spielen, wie Inzidenzmatrizen und Inzidenzlisten. [Kne19; Die00]

2.1 Grundlegende Metriken

Ein einem ersten Teil sollen grundlegende Graph-Kennzahlen vorgestellt werden. Diese beschreiben einen Graphen auf rudimentäre Art und Weise und zeigen die am einfachsten zu berechnenden Eigenschaften des Graphen.

Ordnung und Größe eines Graphen

Die Frage danach, wie viele Knoten ein Graph hat lässt sich mit der **„Ordnung“** eines Graphen beantworten. Die Ördnung”beschreibt dabei einfach die Anzahl der Elemente in der Menge V. Man schreibt: $|V|\ oder\ |V(G)|\ oder\ auch\ |G|$. [Die00] Diese Eigenschaft ist essentiell zur allgemeinen Beschreibung und z.B. graphischen Darstellung eines Graphen. Sie lässt sich dabei in sämtlichen mathematischen Bibliotheken finden, wie SageMath, Matlab und Wolfram [Sag20; Mat20; Wol20]. Die Komplexität zur Erfassung der Metrik gestaltet sich dabei äußerst einfach. Bei einer Adjazenzmatrix lässt sich die Anzahl der Knoten dadurch herausfinden, wie „lang" eine Dimension des zweidimensionalen Arrays bzw. der zweidimensionalen Liste. Dies kann man je nach Implementierung der jeweiligen Datenstruktur in einer Komplexität von $O(n)$ oder $O(1)$ herausfinden.

Eine weitere grundlegende Kennzahl von Graphen ist dessen **„Größe"**. Die Größe beschreibt dabei die Anzahl der Kanten, die im Graphen vorkommen, also die Anzahl der Elemente in der Menge E. Man schreibt analog zur Größe des Graphen: $|E|\ oder\ |E(G)|\ oder\ auch\ ||G||$. [Bal97; Die00] Auch diese Metrik ist weit verbreitet. So lässt sie sich in vielen Büchern zur Graphentheorie finden, aber auch in den genannten Mathematikbibliotheken [Sag20; Mat20; Wol20]. Die Anzahl der Kanten innerhalb eines Graphen herauszufinden, erweist sich nicht ganz so trivial wie das Herausfinden der Ordnung. Ist ein Graph nicht gerichtet, d.h. seine Kanten haben keine feste Richtung [Die00] so ist seine Adjazenzmatrix symmetrisch. Man kann also zählen wie viele Einträge es innerhalb der Matrix auf der Hauptdiagonalen und einer der Hälften gibt. Das ergäbe immer $\frac{1}{2}n^2$ Schritte, wenn n die Ordnung des Graphen ist. Die Komplexität beträge also $O(n^2)$. Bei der Darstellung durch eine Inzidenzliste wäre das anders. Hier könnte einfach die Größe der Liste gesucht werden und man wüsste die Größe des Graphen. Die Komplexität wäre hier, wie schon erwähnt, je nach Implementierung $O(n)$ oder $O(1)$.

Der Grad eines Knotens

Während die zwei ersten vorgestellten Metriken vor allem den Graphen als ganzes beschreiben, ist es auch noch wichtig zu wissen, was einen einzelnen Knoten auszeichnet, um einen Graphen besser zu beschreiben. Hierzu gibt es die grundlegende Metrik des **Grad** eines Knotens. Der Grad eines Knotens beschreibt die Anzahl der mit einem Knoten inzidenten Kanten [Die00]. D.h. es drückt aus, wie viele Kanten mit einem Knoten verbunden sind. Man kann dies z.B. durch eine Funktion ausdrücken, die einen Knoten v auf eine natürliche Zahl abbildet: $d(v)$.

Auf Basis dieser Metrik lässt sich auch andere verwandte Metriken ableiten. Hierzu gehört der **„Minimalgrad"** und der **„Maximalgrad"**. Der Minimalgrad ist der kleinste Knoten-Grad eines Graphen G: $\delta(G) := min\{d(v) \mid v \in V\}$. Parallel dazu ist der Maximalgrad der größte Knoten-Grad eines Graphen G: $\Delta(G) := max\{d(v) \mid v \in V\}$. Darüber hinaus kann man noch den **„Durchschnittsgrad"** eines Graphen bestimmen. Dieser bildet den Durchschnitt aller Knotengrade ab: $d(G) := \sum_{v \in V} d(v)/|V|$. [Die00]

Des Weiteren gibt es bei der Betrachtung eines gerichteten Graphen zusätzliche

Abwandlungen der Metrik. Da hier die Kanten immer zu einem Knoten gerichtet sind unterscheidet man speziell zwischen dem **„Eingangsgrad“** und dem **„Ausgangsgrad“**. Der Eingangsgrad eines Knoten beschreibt dabei die Anzahl der Kanten, die auf einen Knoten „zeigen“. Der Ausgangsgrad zeigt wie viele Kanten von einem Knoten „weggehen“. [Bal97]

Auch der Grad eines Knotens und die meisten seiner verwandten Metriken sind weit verbreitet. So sind der allgemeine Grad, der Eingangsgrad, der Ausgangsgrad in allen drei betrachteten Mathematikbibliotheken vorhanden. SageMath unterstützt sogar nativ die Metrik „Durchschnittsgrad“. [Sag20; Mat20; Wol20]

Die Berechnung eines Grades über eine Adjazenzmatrix oder eine Adjazenzliste ist in linearer Zeit lösbar ($O(n)$). Bei der Adjazenzmatrix muss einfach nur die jeweilige Reihe des zugehörigen Knotens durchlaufen werden und gezählt werden, wie häufig ein Eintrag für eine Kante enthalten. Mit Hilfe der Adjazenzliste kann einfach die Größe der Liste als Grad genommen werden, die dem Knoten zugehörig ist.

2.2 Distanz-Metriken

2.3 Zusammenhangsmetriken (Connectivity)

2.4 Zentralitätsmetriken

2.5 Chromatische Zahl und chromatischer Index

2.6 Weitere Metriken

2.7 Übersicht der vorgestellten Graphmetriken

3 Ähnlichkeit von Graphen

4 Implementierung und Umsetzung der Metriken

4.1 Implementierung in verschiedenen Graphdatenbanken

4.2 Vergleich der Implementierungen

5 Graphmetriken und Ähnlichkeit in Anwendung

6 Fazit und Zusammenfassung

6.1 Zusammenfassung der Ergebnisse

6.2 Fazit

Glossar

Literatur

[Aig15] Martin Aigner. *Graphentheorie: eine Einführung aus dem 4-Farben Problem*. 2., überarbeitete Auflage. Springer Studium Mathematik Bachelor. OCLC: 927721160. Wiesbaden: Springer Spektrum, 2015. 196 S. ISBN: 978-3-658-10322-4 978-3-658-10323-1.

[Alo88] N. Alon. "The linear arboricity of graphs". In: *Israel Journal of Mathematics* 62.3 (Okt. 1988), S. 311–325. ISSN: 0021-2172, 1565-8511. DOI: 10.1007/BF02783300. URL: http://link.springer.com/10.1007/BF02783300 (besucht am 24. 10. 2020).

[And77] Lars Dovling Andersen. "On edge-colorings of graphs." In: *MATHEMATICA SCANDINAVICA* 40 (1. Dez. 1977), S. 161. ISSN: 1903-1807, 0025-5521. DOI: 10.7146/math.scand.a-11685. URL: http://www.mscand.dk/article/view/11685 (besucht am 24. 10. 2020).

[Bal97] V. K. Balakrishnan. *Schaum's outline of theory and problems of graph theory*. Schaum's outline series. New York: McGraw-Hill, 1997. 293 S. ISBN: 978-0-07-005489-9.

[Ber01] Claude Berge. *The theory of graphs*. Dover books on mathematics. Mineola, N.Y: Dover, 2001. 247 S. ISBN: 978-0-486-41975-6.

[BI19] José Bento und Stratis Ioannidis. "A family of tractable graph metrics". In: *Applied Network Science* 4.1 (Dez. 2019), S. 107. ISSN: 2364-8228. DOI: 10.1007/s41109-019-0219-z. URL: https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0219-z (besucht am 22. 10. 2020).

[BK79] Frank Bernhart und Paul C Kainen. "The book thickness of a graph". In: *Journal of Combinatorial Theory, Series B* 27.3 (Dez. 1979), S. 320–331. ISSN: 00958956. DOI: 10.1016/0095-8956(79)90021-2. URL: https://linkinghub.elsevier.com/retrieve/pii/0095895679900212 (besucht am 24. 10. 2020).

[BM08] Raymond Bisdorf und Jean-Luc Marichal. *Counting non-isomorphic maximal independent setsof then-cycle graph*. Nov. 2008. URL: https://arxiv.org/abs/math/0701647v2 (besucht am 24. 10. 2020).

[Die00] Reinhard Diestel. *Graphentheorie*. 2., neu bearb. und erw. Aufl. Springer-Lehrbuch. OCLC: 247312585. Berlin: Springer, 2000. 314 S. ISBN: 978-3-540-67656-0.

[EK13] W. Ellens und R. E. Kooij. *Graph measures and network robustness*. _eprint: 1311.5064. 2013.

[Gus83] Dan Gusfield. "Connectivity and edge-disjoint spanning trees". In: *Information Processing Letters* 16.2 (Feb. 1983), S. 87–89. ISSN: 00200190. DOI: 10.1016/0020-0190(83)90031-5. URL: https://linkinghub.elsevier.com/retrieve/pii/0020019083900315 (besucht am 24. 10. 2020).

[HM11] Javier Martın Hernandez und Piet Van Mieghem. *Classification of graph metrics*. 2011. URL: https://www.nas.ewi.tudelft.nl/people/Piet/papers/TUDreport20111111_MetricList.pdf (besucht am 22. 10. 2020).

[Hun14] Michael Hunger. *Neo4j 2.0 Eine Graphdatenbank für alle*. OCLC: 875609599. 2014. ISBN: 978-3-86802-315-2 978-3-86802-654-2. URL: https://nbn-resolving.org/urn:nbn:de:101:1-2014032620976 (besucht am 24. 10. 2020).

[Jan20] JanusGraph. *JanusGraph Dokumentation*. 2020. URL: https://docs.janusgraph.org/ (besucht am 25. 10. 2020).

[Kne19] Helmut Knebl. *Algorithmen und Datenstrukturen: Grundlagen und probabilistische Methoden für den Entwurf und die Analyse*. OCLC: 1123167896. 2019. ISBN: 978-3-658-26511-3.

[Lov12] László Lovász. *Large networks and graph limits*. American Mathematical Society colloquium publications volume 60. Providence, Rhode Island: American Mathematical Society, 2012. 475 S. ISBN: 978-0-8218-9085-1.

[Mat20] Matlab. *Directed and Undirected Graphs - MATLAB & Simulink - MathWorks Deutschland*. Directed and Undirected Graphs. 2020. URL: https://de.mathworks.com/help/matlab/math/directed-and-undirected-graphs.html (besucht am 10. 11. 2020).

[Moh89] Bojan Mohar. "Isoperimetric numbers of graphs". In: *Journal of Combinatorial Theory, Series B* 47.3 (Dez. 1989), S. 274–291. ISSN: 00958956. DOI: 10.1016/0095-8956(89)90029-4. URL: https://linkinghub.elsevier.com/retrieve/pii/0095895689900294 (besucht am 24. 10. 2020).

[Öst02] Patric R.J. Östergård. "A fast algorithm for the maximum clique problem". In: *Discrete Applied Mathematics* 120.1 (Aug. 2002), S. 197–207. ISSN: 0166218X. DOI: 10.1016/S0166-218X(01)00290-6. URL: https://linkinghub.elsevier.com/retrieve/pii/S0166218X01002906 (besucht am 24. 10. 2020).

[RWE15] Ian Robinson, Jim Webber und Emil Eifrem. *Graph databases [new opportunities for connected data*. OCLC: 1028626678. Sebastopol, CA: O'Reilly, 2015. ISBN: 978-1-4919-3200-1. URL: https://neo4j.com/graph-databases-book/?ref=home (besucht am 24. 10. 2020).

[Sag] SageMath. *Generic graphs (common to directed/undirected) — Sage 9.2 Reference Manual: Graph Theory*. Sage Math Reference Manual. URL: https://doc.sagemath.org/html/en/reference/graphs/sage/graphs/generic_graph.html (besucht am 10. 11. 2020).

[Sag20] SageMath. *Graph Theory*. Sage Math Reference Manual. 2020. URL: https://doc.sagemath.org/html/en/reference/graphs/index.html (besucht am 25. 10. 2020).

[SU11] Edward R. Scheinerman und Daniel H. Ullman. *Fractional graph theory: a rational approach to the theory of graphs*. Dover books on matehmatics. OCLC: ocn721885660. Minola, N.Y: Dover Publications, 2011. 211 S. ISBN: 978-0-486-48593-5.

[Tig20] TigerGraph. *TigerGraph Documentation*. 2020. URL: https://docs.tigergraph.com/ (besucht am 25. 10. 2020).

[Tit19] Peter Tittmann. *Graphentheorie: Eine anwendungsorientierte Einführung*. 3., aktualisierte Auflage. München: Hanser, Carl, 2019. 168 S. ISBN: 978-3-446-46052-2 978-3-446-46503-9.

[Tur04] Volker Turau. *Algorithmische Graphentheorie*. Oldenbourg Wissenschaftsverlag, 1. Jan. 2004. ISBN: 978-3-486-59377-8. DOI: 10.1524/9783486593778. URL: https://www.degruyter.com/view/title/310250 (besucht am 24. 10. 2020).

[WM19] Peter Wills und Francois G. Meyer. *Metrics for Graph Comparison: A Practitioner's Guide*. _eprint: 1904.07414. 2019.

[Wol20] Wolfram. *Graph Measures & Metrics*. Wolfram Language Documentation. 2020. URL: https://reference.wolfram.com/language/guide/GraphMeasures.html (besucht am 25. 10. 2020).