

Interaktive Visualisierung zur Darstellung und Bewertung von Learning-Analytics-Ergebnissen in Foren mit vielen Teilnehmern

Marcus Klüsener¹, Wojciech Konitzer² und Albrecht Fortenbacher³

Abstract: Ausgehend von klassischen Dashboards, geht die Entwicklung bei der Darstellung von Learning-Analytics-Ergebnissen hin zu interaktiven Visualisierung, welche die Interpretation und Anwendung der Ergebnisse des maschinellen Lernens oder der Sozialen-Netzwerk-Analyse unterstützen. In der vorliegenden Arbeit werden zwei visuelle Analysen vorgestellt, welche für das Learning-Analytics-Tool LEMO entwickelt wurden, und auf Iversity MOOCs sowie das Diskussionsforum Stack Overflow angewendet wurden.

Keywords: Learning Analytics, Machine Learning, Visual Analytics, Social Network Analysis.

1 Einleitung

Learning-Analytics-Tools unterstützen Dozent/inn/en und Bildungsanbieter dabei, Lernvorgänge besser zu verstehen und unter Umständen den Lernerfolg vorherzusagen. Mit dem zunehmenden Erfolg von Massive Open Online Courses (MOOCs) steigen auch die Anforderungen an Educational Data Mining, was Methoden und Visualisierung angeht. Dies ist zum einen in der großen Zahl der Studierenden begründet, welche hohe Anforderungen an die Visualisierung von Lernvorgängen stellt, zum anderen stellen veränderte Lernverhalten, wie man sie aus einer Completion Rate von unter 5% in MOOCs vermuten kann, beispielsweise hohe Anforderungen an verwendete Klassifikationsverfahren.

Die beiden vorgestellten Entwürfe für die Visualisierung von Klassifikationsergebnissen, und für die Visualisierung von großen Lerner-Netzwerken, entstanden im Rahmen des LEMO-Projekts [FEM13], in Kooperation mit dem MOOC-Plattform-Betreiber Iversity. In der ersten Analyse werden Klassifikationsergebnisse, welche eine Vorhersage der Performance von Studierenden erlauben, mit Hilfe von Scatterplots visualisiert. Durch die interaktive Erkundung der Korrelation von Merkmalskombinationen mit der Klassifikation der Teilnehmer (nach ihrer Performance) können nicht nur Zusammenhänge veranschaulicht, sondern auch die verschiedenen eingesetzten Klassifikationsverfahren

¹ HTW Berlin, Wilhelminenhofstraße 75A, 12459 Berlin, kluesen@htw-berlin.de

² Fraunhofer Institut für Offene Kommunikationssysteme (FOKUS), Kaiserin-Augusta-Allee 31, 10589 Berlin, wojciech.konitzer@fokus.fraunhofer.de

³ HTW Berlin, Wilhelminenhofstraße 75A, 12459 Berlin, forte@htw-berlin.de

überprüft und Werte validiert werden.

Die zweite visuelle Analyse stellt Merkmale der Lernenden, welche als Daten in Diskussionsforen vorhanden sind, oder aus den Daten abgeleitet werden können, dar. Das Problem hierbei ist, Ausschnitte aus dem Netzwerk-Graphen darzustellen, welche Strukturen erkennen lassen. Eine wesentliche Rolle spielt hierbei der Begriff der Zentralität.

Zwei wesentliche Features des Tools LEMO sind die Unabhängigkeit von der Plattform, auf welcher die Aktivitätsdaten der Lernenden gesammelt werden, und eine interaktive intuitive Visualisierung der Analyseergebnisse [Be13]. Die Plattformunabhängigkeit wird durch die Übernahme der Daten in ein LEMO-Datenmodell erreicht [FKS14]. Dadurch war es möglich, Analysen und Visualisierungen für verschiedene Lernumgebungen, von Moodle-Kursen über sehr große MOOCs bis hin zu dem Diskussionsforum Stack Overflow, welches eine stark frequentierte Plattform für Fragen aus dem Bereich der Informatik ist, zu unterstützen.

2 Erkundung höherdimensionaler Merkmalsräume mit Hilfe von Scatterplots

In [K115] wurde zur Vorhersage der Studienleistung in Iversity MOOCs ein Klassifikator entwickelt, welcher auf Aktivitätsdaten in den MOOC-Foren basiert. Das Ergebnis des Klassifikators wurde zur Verdeutlichung des Volumenflusses als Sankey Diagramm (siehe Abbildung 1) dargestellt. Um die Ergebnisse des Klassifikators zu bewerten, wurde eine grafische Darstellung mit einem Scatterplot (dt. Streudiagramm) gewählt.

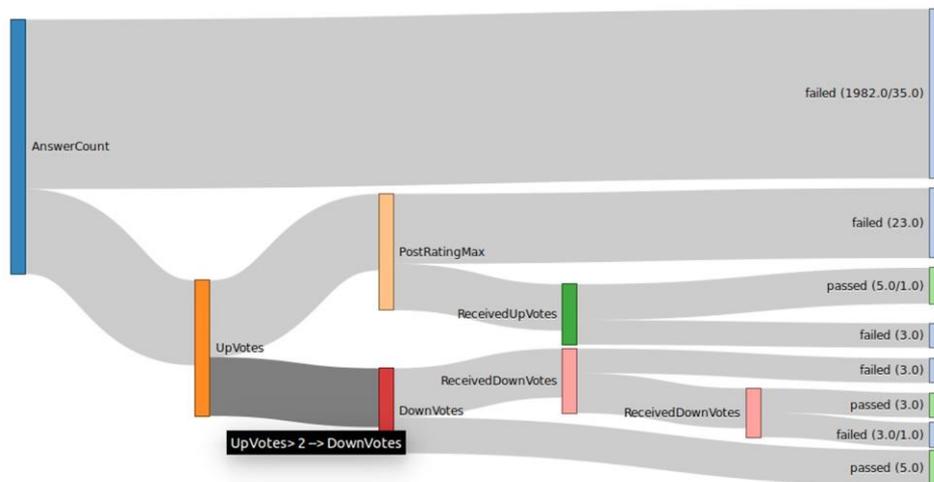


Abb. 1: Sankey-Diagramm zur Beschreibung des Volumenflusses

Bei Scatterplots können unterschiedliche Merkmale für die Darstellung auf der x- und y-Achse ausgewählt werden, und interaktiv alle Merkmale eines Datensatzes untersucht werden [Ve88]. In der Anordnung der Instanzen des Datensatzes können sich Muster ergeben, die Informationen über die Abhängigkeitsstruktur der beiden Merkmale erkennen lassen. Bei komplexeren Datensätzen und zunehmender Anzahl an darzustellenden Variablen wird die Interpretierbarkeit jedoch schwieriger. Sedlmair et al. haben empirisch untersucht, wie Probanden die Klassen in 2D Scatterplots, interaktiven 3D Scatterplots und Scatterplot-Matrizen (SPLOMs) mit Dimensionsreduzierung trennen konnten. Sie kamen zu dem Fazit, dass 2D Scatterplots ausreichend zur Separierung der Klassen seien [SMT13].

Ein Dozent kann mit dem Scatterplot eine explorative Analyse durchführen, durch Auswertung aller möglichen Merkmalskombinationen, oder durch Einschränkung auf Kombinationen von für die Klassifikation wichtigen Merkmalen (siehe Abbildung 2). Dabei können Hypothesen über die Ursache und den Grund der beobachteten Daten gebildet werden [He15]. Eine Gruppenbildung der Teilnehmer, welche die untersuchten Klassen separiert, deutet darauf hin, dass die gewählte Merkmalskombination gut für die Klassifikation geeignet ist.

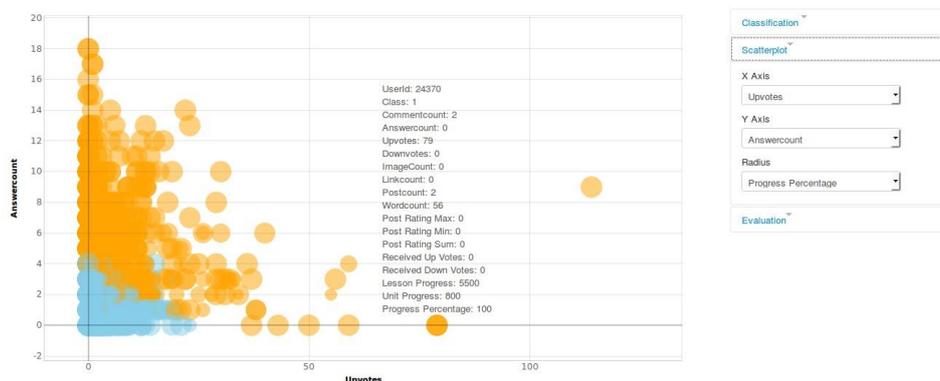


Abb. 2: Scatterplot zur Darstellung der erfolgreichen Studierenden (blau) und der nicht so erfolgreichen Studierenden (gelb)

In dem LEMO-Scatterplot werden die Klassen der Studierenden farblich unterschiedlich dargestellt, wobei das Problem die große Anzahl der Instanzen darstellt. Die Lösung dafür besteht in einer Darstellung einer Instanz (eines Studenten) durch einen kleinen Kreis, welcher transparent mit der Klassenfarbe gefüllt wird. Viele Instanzen, welche nahe zusammenliegen, können als intensiver Farbton erkannt werden. Je besser die Klassen in einer gewählten Merkmalskombination separiert werden, desto deutlicher sind die Farben zu erkennen, Mischfarben deuten auf eine ungenügende Separation hin.

Durch die Kombination der beiden Visualisierungen kann ein Dozent jetzt visuell den

Klassifikator verifizieren. Er kann nachvollziehen, dass diejenigen Merkmale, die laut Sankey-Diagramm für die Klassifikation besonders relevant sind, die im Scatterplot dargestellten Instanzen am besten in Gruppen teilen. Klassifikationsfehler können durch das in der Größe der Punkte dargestellte Klassenattribut (Performance) erkannt und bewertet werden. Ein als bestehend klassifizierter Student mit einem kleinen Radius kann als falsch klassifiziert identifiziert werden, falls die Performance des Studenten unter dem Schwellwert zum Bestehen des Kurses liegt. Über einen Tooltip können die Merkmale angezeigt werden, und ein Dozent hat die Möglichkeit, mit zusätzlichem Wissen über diesen Studenten den Klassifikationsfehler zu bewerten. Wird der Klassifikator auf einen laufenden Kurs angewendet, kann der Dozent an Hand der farblichen Einteilung erkennen, welche Studenten laut Klassifikation vermutlich bestehen werden.

Die im Scatterplot und im Sankey Diagramm dargestellten Merkmale wurden als Indikatoren für Forenbeteiligung aus der Datenbasis extrahiert. Die Merkmale konnten zum Teil direkt aus den Daten abgeleitet werden (Upvotes / Downvotes) oder wurden durch Analysen berechnet (Anzahl der Bilder). Im Scatterplot in Abbildung 2 sind die laut Sankey-Diagramm relevantesten Merkmale Answercount und Upvotes aufgetragen. Die Farbe entspricht dem Klassifikationsergebnis und die Größe dem Klassenattribut. Über den Tooltip hat der Nutzer die Möglichkeit, alle verfügbaren Merkmale anzuzeigen.

3 Darstellung von Merkmalen ausgewählter Studierender in großen Netzwerken

In [Ko15] wurde auf Basis von Forenaten aus der Iversity MOOC-Plattform und des Diskussionsforums Stack Overflow eine soziale Netzwerkanalyse (SNA) durchgeführt, um die soziale Netzwerkstruktur der Studierenden und deren Aktivitäten im Lernnetzwerk abzubilden und verständlich zu machen. Die visuelle Analyse wird in das Tool LEMO integriert.

Die Visualisierung basiert auf großen Mengen von Lernerdaten, wie sie auf Lernplattformen, in sozialen Netzen oder Diskussionsforen gesammelt werden, und dient dem besseren Verständnis komplexer Lernvorgänge. Der Entwicklung lagen folgende Fragestellungen zu Grunde [Sc13]:

- Wer spricht mit wem?
- Zu welchem Kursthema entstehen Lerngemeinschaften, und wie lassen sich diese darstellen?
- Wie entstehen Beziehungen zwischen den Akteuren in einem Forum?
- Wie intensiv sind die Beziehungen zwischen den Nutzern?
- Wer ist ein Experte oder nur Konsument von Wissen?

- Wer ist am aktivsten? Wie aktiv ist ein Netzwerk?
- In welchen Kursen haben die meisten Studenten Verständnisprobleme?

Die realisierte interaktive Visualisierung soll einige dieser Fragen beantworten helfen, und es dem Dozenten ermöglichen, seine Lehre an die Bedürfnisse der Studierenden anzupassen.

Die SNA vereint Analyseverfahren und Informationsvisualisierung. In einer formalen Analyse werden die sozialen Beziehungsstrukturen und das soziale Verhalten von Akteuren in Foren untersucht. Die Akteure betreiben einen Informationsaustausch und verfügen über Gemeinsamkeiten. Das Netzwerk wird mit Hilfe eines Soziogramms abgebildet, das die Nutzer als Knoten und die Beziehungen zwischen den Nutzern als Kanten darstellt. Die Wahl und Festlegung von Darstellungsmerkmalen ausgewählter Studierender in einem Netzwerk beschränkt sich dabei nicht nur auf einfarbige Netzwerkknoten, sondern jeder Knoten erhält weitere wichtige, visuelle Informationen über die Nutzeraktivität.

Um möglichst viele Informationen über einen Akteur im Netzwerk zu erhalten, wurde als Darstellungsform für einen Knoten ein Kreisdiagramm gewählt. Das Kreisdiagramm besteht aus maximal fünf Teilwerten (siehe Abbildung 3). Dazu zählen die Anzahl der gestellten Fragen, geschriebenen Antworten und Kommentare auf eigene und fremde Beiträge zu einem bestimmten Kurskapitel im Forum. Außerdem besitzt ein Knoten Informationen über den letzten Aktivitätszeitpunkt. Gerichtete Kanten geben Informationen über die Kommunikationsrichtung und lassen erkennen, von welchem Nutzer das Handeln ausgeht.

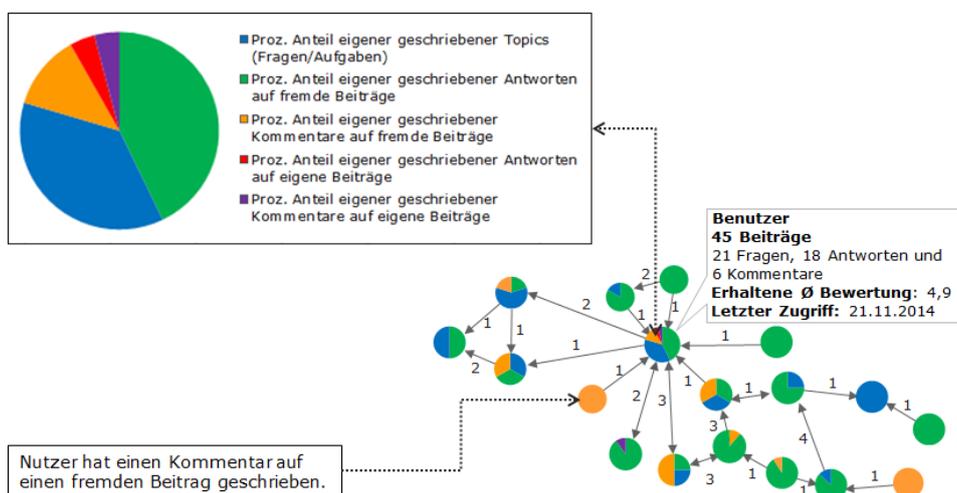


Abb. 3: Entwurf eines Forenaktivitätsgraphen für LEMO

Die interaktive Visualisierung gibt dem Anwender die Möglichkeiten der Lernanalyse und bringt neues Wissen hervor. Zusätzliche Filter erlauben es dem Anwender, die zu visualisierende Datenmenge einzuschränken. Einer dieser Filter erlaubt die Auswahl bestimmter Nutzer. Es können der aktivste Forennutzer, die 10 aktivsten Forennutzer oder bekannte aktive Nutzer ausgewählt werden. Abhängig von dieser Auswahl entstehen verschiedene Nutzernetzwerke (siehe Abbildung 5). Die SNA verfügt über verschiedene Verfahren zur Analyse eines sozialen Netzwerks und der Akteure eines Netzwerks. Dazu zählen u.a. die Netzwerkdichte (engl. density), Extraktion von Lerngemeinschaften und Identifizierung von besonders engagierten Akteuren (Experten) innerhalb eines Netzwerks. Die Netzwerkdichte (siehe Abbildung 4) beschreibt, wie intensiv die Kommunikation in einem Netzwerk ist. Ein Netzwerk, in dem jeder Knoten mit jedem anderen Knoten über Kanten verbunden ist, entspricht einer Dichte von 100%. Die berechnete Netzwerkdichte wird oberhalb des visualisierten Netzwerkgraphen ausgegeben.

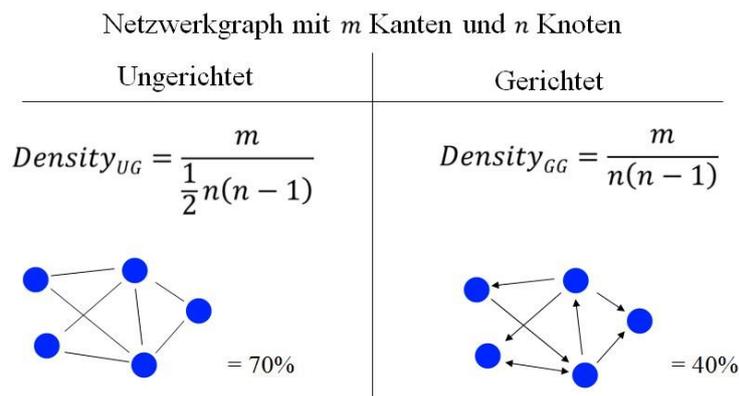


Abb. 4: Berechnung der Netzwerkdichte

Mit Hilfe der Zentralitätsberechnung kann bestimmt werden, welche die wichtigsten und aktivsten Akteure in einem sozialen Netzwerk sind. So kann beispielsweise mit der Zwischenzentralität berechnet werden, über welchen Knoten die meisten kürzesten Pfade in einem Netzwerk verlaufen. Ein Knoten mit einer hohen Zwischenzentralität befindet sich auf einem hohen Anteil von kürzesten Pfaden und verbindet häufig zwei Teilnetze zu einem Netzwerk. Er dient als Kommunikationsvermittler. Abhängig der gewählten Zentralitätsart der berechneten Zentralitäten verändern sich alle Knotenradien interaktiv im dargestellten Netzwerk. Je wichtiger ein Akteur ist, desto größer ist sein Knoten. Statt der errechneten Zentralität kann auch die Anzahl geschriebener Beiträge im Forum eines Nutzers zur Knotengröße berücksichtigt werden. Je mehr Beiträge verfasst wurden, desto aktiver ist ein Nutzer.

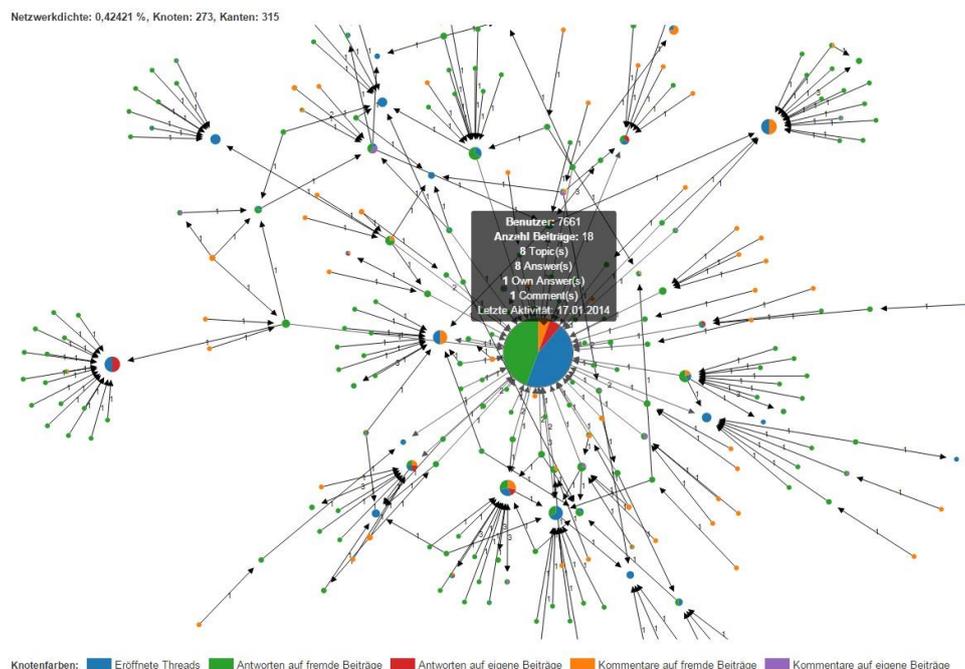


Abb. 5: Netzwerkausschnitt des aktivsten Benutzers mit ID 7661 zu einem bestimmten Thema im Forum, gradbasierte Zentralität (die Knotengröße ist abhängig von der Anzahl direkter Kanten)

4 Bewertung und Ausblick

Beide vorgestellten Visualisierungen bieten interessante Einblicke in das Lernverhalten sehr vieler Teilnehmer, basierend auf ihrem Kommunikationsverhalten in Diskussionsforen. Im ersten Fall war es das Q&A-Forum in Iversity MOOCs, welches primär die Kommunikation Dozent/in und Lernende abbildet. Im zweiten Fall (Visualisierung des Kommunikationsverhaltens auf der Plattform Stack Overflow) war dagegen der Begriff Zentralität, bzw. Zwischenzentralität, ein geeignetes Merkmal, um beispielsweise aktive Studierende oder Strukturen des Lernenden-Netzwerks zu identifizieren.

Bei der Visualisierung der Klassifikationsergebnisse zur Performance der Lernenden mit Hilfe von Scatterplots und Sankey Diagrammen konnten nicht nur die Ergebnisse verifiziert, sondern auch nicht-offensichtliche Zusammenhänge zwischen Merkmalen und Performance(-Vorhersage) erkannt werden. Durch Verbesserungen der Klassifikation, beispielsweise durch Einschränkung der Datenbasis, oder durch eine differenzierte Klasseneinteilung, können auch bessere Visualisierungsergebnisse erwartet werden [KPS13].

Die neue Versions LEMO2, welche gegenwärtig in Entwicklung ist, soll durch eine veränderte Systemarchitektur, ein neues Datenmodell und durch eine neue Analysestruktur obige Analysen und Visualisierungen umsetzen, und insbesondere den Anforderungen an Learning Analytics in Lernumgebungen mit sehr vielen Teilnehmern (MOOCs) Rechnung tragen.

Literaturverzeichnis

- [Be13] Beuster, Liane; Elkina, Margarita; Fortenbacher, Albrecht; Kappe, Leonard; Merceron, Agathe; Pursian, Andreas; Schwarzrock, Sebastian; Wenzlaff, Boris: Learning Analytics und Visualisierung mit dem LeMo-Tool. In: DeLFI. S. 245–250, 2013.
- [FEM13] Fortenbacher, Albrecht; Elkina, Margarita; Merceron, Agathe: The Learning Analytics Application LeMo–Rationals and First Results. *International Journal of Computing*, 12(3):226–234, 2013.
- [FKS14] Fortenbacher, Albrecht; Klüsener, Marcus; Schwarzrock, Sebastian: Ein generisches Datenmodell für Learning Analytics. In: C. Rensing & S. Trahasch, *Proceedings of DeLFI Workshops*. 2014.
- [He15] He, Jiazhen; Bailey, James; Rubinstein, Benjamin IP; Zhang, Rui: Identifying At-Risk Students in Massive Open Online Courses. 2015.
- [KI15] Klüsener, Marcus: Vorhersage der Studienleistung durch Forenanalyse und Klassifikationsverfahren im Learning-Analytics-Tool LEMO. Masterarbeit, Hochschule für Technik und Wirtschaft Berlin, 2015.
- [Ko15] Konitzer, Wojciech: Learning Analytics und Interaktive Visualisierung von Social Media Daten. Masterarbeit, Hochschule für Technik und Wirtschaft Berlin, 2015.
- [KPS13] Kizilcec, René F; Piech, Chris; Schneider, Emily: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: *Proceedings of the third international conference on learning analytics and knowledge*. ACM, S. 170–179, 2013.
- [Sc13] Schön, Martin; Ebner, Martin et al.: Das Gesammelte interpretieren. *Educational Data Mining und Learning Analytics*. In: L3T. Lehrbuch für Lernen und Lehren mit Technologien. 2013.
- [SMT13] Sedlmair, Michael; Munzner, Tamara; Tory, Melanie: Empirical guidance on scatterplot and dimension reduction technique choices. *Visualization and Computer Graphics*, IEEE Transactions on, 19(12):2634–2643, 2013.
- [Ve88] Verboon, Peter: *Graphical Tools in Multivariate Analysis*. Department of Data Theory, University of Leiden, 1988.