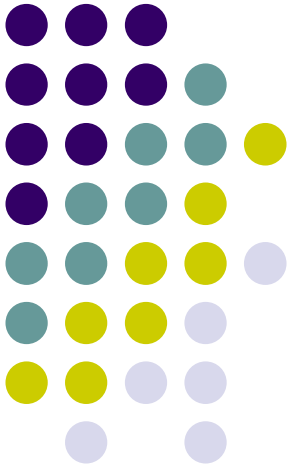
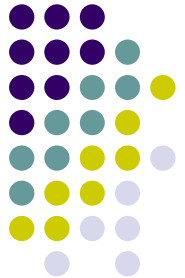
Statistik in der Soziologie, 1950 - 2000



Angela Bosnjak

29.11.11

Gliederung



- 0. Einführung

- 1. **Die erste Generation: Kreuztabellen**
- 1.1. Kategoriale Datenanalyse
- 1.2. Signifikanztest und Modellselektion

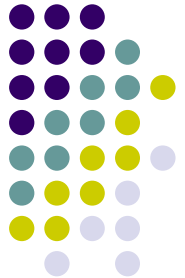
- 2. **Die zweite Generation: Unit-level Daten**
- 2.1. Messverfahren des Berufsstatus
- 2.2. Anwendungen der Strukturgleichungsmodelle
- 2.3. Ereigniszeitanalyse

- 3. **Die dritte Generation: Neue Datensätze, neue Herausforderungen, neue Methoden**
- 3.1. Soziale Netzwerke und Daten der 3. Generation
- 3.2. Analyse von Textdaten
- 3.3. Narrative und sequenzielle Analyse
- 3.4. Simulationsmodelle
- 3.5. Makrosoziologie

- 4. **Soziale Netzwerke und soziale Netzwerkanalyse**
- 4.1. Definitionen und Erläuterungen
- 4.2. Methoden zur Visualisierung

- 5. Diskussion

0. Einführung



Kurzer Überblick:

- Mitte der 90-er: Ursprung der Soziologie mit der industriellen Revolution
- Vor dem 2. WK: Daten sehr lückenhaft und ungenau
- Nach dem 2. WK: Gewinnung von komplexen Daten und somit auch Entwicklung statistischer Methoden

- Einteilung in 3 Generationen:
  - seit dem 2. WK: Kreuztabellen; kleine Anzahl an diskreten Variablen
  - seit den frühen 60er Jahren: Unit-level Daten mit vielen Variablen
  - späte 80er Jahre: Daten bekommen verschiedene Formen; Abhängigkeit der Daten im Vordergrund

- Heute: neue Darlegung von großen, hochwertigen Stichprobenerhebungen (Umfang: 5.000 - 20.000)

1. Die erste Generation: Kreuztabellen

1.1. Kategoriale Datenanalyse

- Großteil der Datensätze in Form von Kreuztabellen
- Kategorienanzahl zwischen 5 und 17

- Birchs Vorschlag: Log-lineares Modell für beobachtete Zählungen $\{x_{ij}\}$

$$\log(E[x_{ij}]) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

- Erfolgreicher: Modellassoziation von Duncan und Goodman:

$$u_{12(ij)} \sum_{k=1}^{K} \gamma_k \alpha_i^{(k)} \beta_j^{(k)} + \phi_i \delta(i, j)$$

- Hout: Erweiterung der Modellierung dieser Auswertungen (Scores) und der diagonalen Bedingungen zu Summen oder Produkten von Kovariaten
- Alternative Formulierung in Form von Randverteilungen und Odds Ratios impliziert das Modell für die gemeinsame Verteilung

Table Überlebt*Geschlecht

			Geschlecht		
			Frauen	Männer	TOTAL
Überlebt	nein	observed	126	1364	1490
		expected	318.17	1171.83	1490
	ja	observed	344	367	711
		expected	151.83	559.17	711
	TOTAL	observed	470	1731	2201
		expected	470	1731	2201

Table 6
Social Mobility of British Males

	Subject's status		
Father's status	Upper	Middle	Lower
Upper	588	395	159
Middle	349	714	447
Lower	114	320	411

Note. Data in this table are from *The Analysis of Contingency Tables* (p. 111) by B. S. Everitt, 1977, London: Chapman & Hall. Copyright 1977 by Chapman & Hall. Reprinted by permission.

1. Die erste Generation: Kreuztabellen

1.2. Signifikanztest und Modellselektion

- Soziologen haben Stichprobengrößen in einer vierstelligen Zahl
- Problem: Die p-Werte indizieren die Ablehnung von Nullhypothesen in großen Stichproben, obwohl das Nullmodell als theoretisch sinnvoll erscheint

- Anfang 80er: Einige Soziologen ignorierten die Ergebnisse der Tests die auf den p-Wert basierten und nutzten die Modellselektion
- Alternative: Bayesianisches Informationskriterium (BIC); Beurteilung der vorliegenden empirischen Daten (Stichprobe) und der Komplexität des Modells gemessen an der Parameteranzahl

$$BIC = -2 \log L_D(M) + d \log n$$

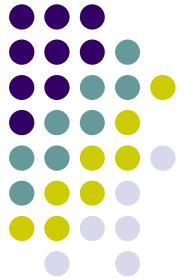
  - Vorteile:
    - Möglichst genaue Theorieprüfung als primäres Forschungsziel
    - Selektiert das sparsamste Modell, das zugleich die K-L Distanz minimiert (-> quasi-wahres Modell) und dessen K-L-Informationsverlust mit steigendem n sinkt

  - Nachteile:
    - Underfitting (bei Spezifizierung eines Modells werden Variablen außer Acht gelassen)
    - Hoher Bias
    - Quasi-wahr ist nicht Wahrheit -> keine Übertragbarkeit auf die Grundgesamtheit

  - Lösung: Berechnung des Bayesfaktors;
    - Darstellung für den Vergleich zweier Hypothesen
    - Anpassung an den klassischen Hypothesentest
    - reflektiert aktuell verfügbare Informationen
    - Entscheidung ohne p-Wert

2. Die zweite Generation: Unit-level Daten

2.1. Messverfahren des Berufsstatus



- Sehr wichtiges Konzept in der Soziologie
- Der Berufsstatus wird gleichgestellt mit dessen jeweiligen Ansehen
- Problem: Mit Hilfe von Umfragen ist nur eine kleine Anzahl von insgesamt 800 Berufen (die im Zensus identifiziert worden sind) messbar

- Duncan stufte die Scores zurück um die fehlenden Werte eintragen zu können (Kategorisierung der Berufe)
- Lösung: Entwicklung des SEI (socioeconomic index) als Prädiktor: Abhängigkeit zwischen Schulbildung, Einkommen und beruflichem Ansehen

- Alternativ: Aktuelles Einkommen oder Gesundheit als Prädiktor; scheitert, da Messungen durch Antwortverweigerung, Widerruf oder Zuverlässigkeit verzerrt wurden

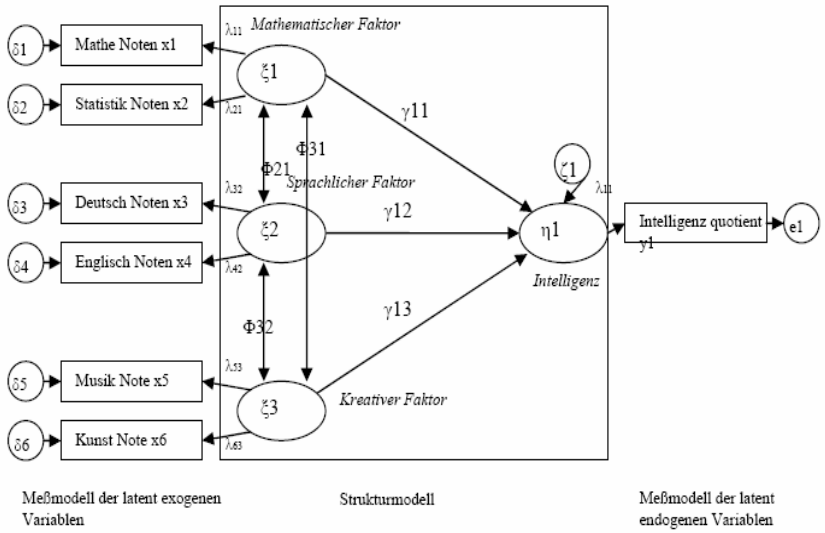
- Berufsstand ist zeitlich am stabilsten und somit auch der beste Einflusswert

2. Die zweite Generation: Unit-level Daten

2.2. Anwendung der Strukturgleichungsmodelle

- Wrights Pfadanalyse: Eine Form der Untersuchung der Abhängigkeiten zwischen Variablen;
  Im Rahmen der Pfadanalyse werden Pfadmodelle, d.h. theoretisch hergeleitete Modelle kausaler Zusammenhänge zwischen Variablen, empirisch überprüft. Die Pfadanalyse ist Teil der Kausalanalyse.

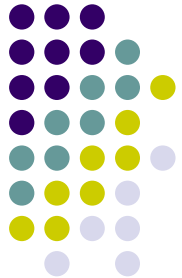
- Interessierende Variablen im Kausalmodell:
  - Wirkungszusammenhang theoretisch belegbar
  - vermutete Abhängigkeit zweier Variablen empirisch nachweisbar
  - Überprüfung von Drittvariablen



- Jöreskogs LISREL-Modell: ML-Schätzung des Strukturgleichungsmodells mit latenten Variablen
- Muthén weitete das LISREL-Modell weiter aus, indem er zunächst kategoriale Variablen verwendete und später damit Längsschnittdaten, Wachstumskurven und multidimensionale Datensätze berechnete

2. Die zweite Generation: Unit-level Daten

2.2. Anwendung der Strukturgleichungsmodelle



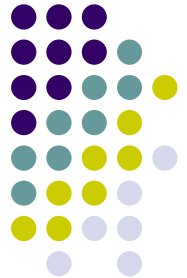
- Gegenstück zum LISREL:
- Markov-Chain-Monte-Carlo-Verfahren (MCMC): Klasse von Algorithmen, die auf Basis der Markov-Kette Stichproben aus den Wahrscheinlichkeitsverteilungen zieht
  - Ziel: Versuch mit Hilfe der Wahrscheinlichkeitstheorie analytisch keine oder nur aufwändig lösbare Probleme numerisch zu lösen auf Basis der Markov-Kette

- Markov-Kette: Bei Kenntnis einer begrenzten Vorgeschichte genauso gute Prognosen über die zukünftige Entwicklung möglich wie bei der Kenntnis der gesamten Vorgeschichte des Prozesses
- Zustand der Kette nach einer großen Zahl von Schritten wird dann als Stichprobe der erwarteten Verteilung benutzt
- Qualität der Stichprobe steigt mit zunehmender Zahl der Schritte

- MCMC-Verfahren wird nun als Erweiterung des LISREL-Modells betrachtet

2. Die zweite Generation: Unit-level Daten

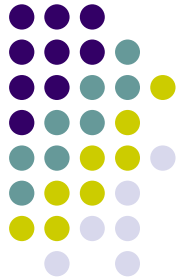
2.3. Ereigniszeitanalyse



- Unit-level Daten erlauben die Rekonstruktion von Lebensgeschichten (bei wichtigen Ereignissen wie z.B. Hochzeit, Scheidung, Geburten, Inhaftierungen, Freilassungen aus dem Gefängnis, Jobwechsel, Sozialhilfebezug, etc.)
- Vergleich der Zeiten bis zu einem bestimmten Ereignis zwischen 2 oder mehreren Gruppen um die Wirkung von prognostischen Faktoren, medizinischen Behandlungen oder schädlichen Einflusses zu schätzen (Mortalität muss vorliegen)

- Beispiel dafür ist die Cox-Regression: Modellierung von Überlebenszeiten in der Survival Analysis basierend auf dem Konzept der Hazardrate

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden



3.1. Soziale Netzwerke und Daten der 3. Generation

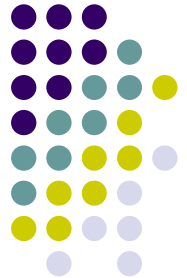
- Soziale Netzwerke bestehen aus einer Menge an paarweisen Verbindungen, wie z.B. Freundschaft zwischen Jugendlichen, sexuelle Beziehungen zwischen Erwachsenen, politische Verbände zwischen sozialen Gruppen, etc.
- Markov-Random-Field (MRF): Statistisches Modell, das ungerichtete Zusammenhänge in einem Feld beschreibt
- Feld: besteht aus Zellen, die Zufallsvariablen enthalten und räumlich begrenzt gegenseitig wechselwirken

- Hammsersley-Clifford-Theorem:
  - gibt erforderliche und ausreichende Konditionen, durch welche eine positive Wahrscheinlichkeitsverteilung als MRF dargestellt werden kann,
  - befasst sich mit den Größen einer Menge von miteinander verbundenen Lokationen, d.h., dass die Größe von dem unmittelbaren Nachbarn abhängt

- Für die stat. Genetik wurden die Stammbaumanalyse benutzt, welche aber nicht so erfolgreich war wie die MCMC-Methode

- Großteil der Daten räumlich (dynamisch); Soziologen ignorierten dies
- Ausnahmen: Masseys und Dentons Studie der örtlichen Rassentrennung und bei ihrer Studie der Geburtenhäufigkeit und Verhütung in Asien

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden



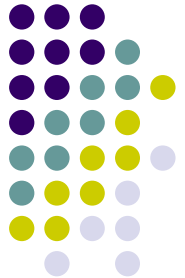
3.2. Analyse von Textdaten

- Die gröbste Form der Auswertungen von soziologischen Daten ist textuell
- Erfolge bei der Formanalyse führten zur Inhaltsanalyse (verschiedenartige Zählungen von Wörtern in Texten)
- Alternative: Viel versprechende, aktuellere Bemühungen den Kontext in Sätzen zu erfassen scheiterten
- Menschlicher Verstand bei weitem erfolgreicher bei der Analyse von Textdaten als der Computer (zumindest bis jetzt)

- Fortschritt: Singer, Ryff, Carr und Magee entdeckten eine verblüffende Betrachtungsweise; sie vermischten quantitative und qualitative Denkansätze, d.h., das standardmäßige, unit-level Datensätze, mit mehr als 250 Variablen pro Person, zu schriftlichen „Biografien“ umgeformt wurden

- Suche nach gemeinsamen Merkmalen in den Biografien der Personen führte zur Verallgemeinerung der Merkmalsbeschreibungen

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden

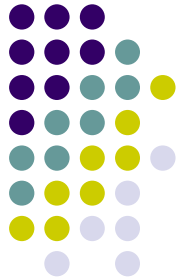


3.3. Narrative und sequenzielle Analyse

- Narrative Analyse: Orientierung an der Analyse des subjektiven Sinns
- Erhebungsverfahren: vorallem narrative Interviews
- Unterscheidung von zwei narrativen Analyseverfahren:
  Analyse narrativer Interviews zur Rekonstruktion von Ereignissen
  Analyse narrativer Daten als Lebenskonstruktion

- Abbott und Hrycak: Einführung der sequenziellen Analyse (Gestalt des Textes gewinnt mehr an Bedeutung)
  - keine Ableitung der Kenntnisse aus Prozessen, die im zeitl. Verlauf des Falles (z.B. Interview) später abgelaufen sind um Unsicherheiten, Mehrdeutigkeiten, etc. der aktuellen Textstelle zu erklären
  - Bedeutungen werden demnach sequentiell aufgeschichtet
  - soll als Ersatzmodell für DNA- oder Proteinsequenzierung darstellen, indem optimale Anpassungsmethoden aus der Molekularbiologie genutzt werden

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden

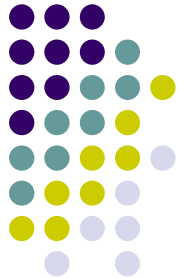


3.3. Narrative und sequenzielle Analyse

- Einsatz von Cluster-Analysen
- Cluster-Analyse: explorative Gruppierung von Objekten nach ihrer Ähnlichkeit bzw. Unähnlichkeit (Distanz)
  - Ziele:
    - eine Gruppe von Objekten zu homogenen Gruppen zusammenzufassen, welche zum Erhebungszeitpunkt noch nicht vorlagen
    - theoriegeleitete Strukturierung der vorliegenden Untersuchungseinheiten um Typologisierungen (wie sozialer Status von Personen oder ökonomische Entwicklungsniveaus von Ländern), auf der Grundlage theoretischer Überlegungen vorzunehmen -> Datenreduzierung

- Abgrenzung: Cluster-Analysen…
  - … erlauben keine inferenzstatistische Prüfung von Gruppierungen bzw. der Wichtigkeit von Merkmalsvariablen für die Gruppierung
  - … beinhalten kein Vorhersagemodell, das für die Prognose von Gruppenzugehörigkeit anderer Personen/Objekte verwendet werden könnte
  - … sind somit keine direkt eigenständigen Verfahren

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden



3.4. Simulationsmodelle

- Zwei Arten: Makro- und Mikrosimulationsmodelle
- Deterministisch, sehr kompliziert und zeigen Systeme in verschieden interagierenden Abteilungen und jede Abteilung in einer Reihe von Differenzialen oder Differenzgleichungen
- Beispiel: Erkundung der Folgen von unterschiedlichen Theorien über den Zusammenhang von der landeseigenen Politik und Krieg oder die Rolle der sexuellen Netzwerke bei der Verbreitung von HIV

- Probleme:
  - Schlechte Einschätzung der involvierten Parameteranzahl
  - Keine Gewissheit, ob es sich um ein passendes Modell handelt
  - Keine Etablierung bei Vergleich mit konkurrierenden Modellen

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden

3.5. Makrosoziologie

- Der Teil der Soziologie, der die Gesellschaft schlechthin zum Gegenstand hat, insofern diese als ein Gefüge von Sozialgebilden begriffen wird
- Makrosoziologie befasst sich mit den Gesetzmäßigkeiten bei der Entwicklung und Veränderung gesellschaftlicher Phänomene, z.B.: Entwicklung eines Volkes, eines gesellschaftlichen Systems oder einer Industrie
- Gegründet auf allgemein vorfindbaren Mustern
- Keine notwendige Abhängigkeit von unmittelbaren Wechselbeziehungen der Mitglieder, wie das bei (Klein-)Gruppen der Fall ist
- Verwendung von sehr großen Einheiten

3. Die dritte Generation: Neue Daten, neue Herausforderungen, neue Methoden

3.5. Makrosoziologie

- Alternative: Bollen und Appolds Konstruktion einer kleineren Stichprobegröße
- Problem: oft nicht möglich; keine allgemeingültige Lösung

- Lösung: Bayes-Schätzung oder Bayes-Inferenz (Anpassen eines Wahrscheinlichkeitsmodells an eine Menge von Daten)
- Bayes-Faktoren…
- …sind in einer kleinen Stichprobe weniger bindend als die Standardsignifikanztests
- … erlauben eine graduierte Begutachtung der Ereignisse anstatt die Ablehnung oder Annahme einer Hypothese zu fokussieren
- … stellen eine Art Buchführung für Modellunsicherheiten bereit, die in diesem Kontext sehr groß sein können

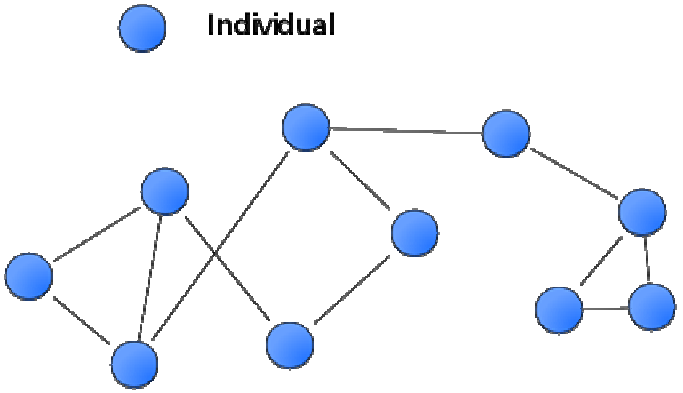
4. Soziale Netzwerke und Soziale Netzwerkanalyse

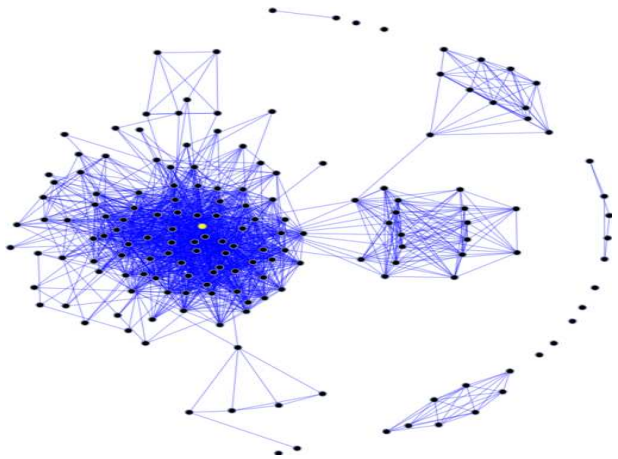
4.1. Definitionen und Erläuterungen

- Soziale Netzwerke lassen sich als Graphen repräsentieren

- Graph G = (V,E)
- …ist eine Menge von Knoten (Vertices) und Kanten (Edges)
- … ist ein formales Modell, dass es erlaubt, die Struktur eines Netzwerkes numerisch in einer Matrix abzubilden und zu analysieren

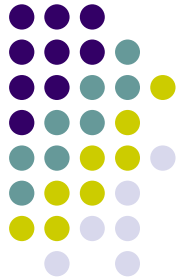
- Vorteil:
  Es besteht die Möglichkeit, dass durch diese formal einfachen Graphendefinitionen aus Knoten und Kanten die unterschiedlichsten Netzwerke mit den gleichen Methoden und Algorithmen zu beschreiben, z.B. Wirtschaftsnetzwerke, Verwandtschaftsnetzwerke, Wissenschaftsnetzwerke, etc.





4. Soziale Netzwerke und Soziale Netzwerkanalyse

4.1. Definitionen und Erläuterungen



- Soziale Netzwerkanalyse: eine Methode der empirischen Sozialforschung zur Erfassung und Analyse sozialer Beziehungen und sozialer Netzwerke.

- Methoden zur Analyse:
  - Verfahren zur Zentralitätsberechnung: Diese zielen darauf ab, die wichtigsten, aktivsten und prominentesten Akteure in einem Netzwerk zu identifizieren.

  - Berechnung von Dichte: Ein Maß zur Charakterisierung von Netzwerken oder Netzwerkteilen ist die Dichte. Sie ist ein Indikator für die gesamte Aktivität eines Netzwerkes. Dichte ist definiert als das Verhältnis der vorhandenen Beziehungen zur Anzahl maximal möglicher Beziehungen. Sie kann einen Wert zwischen 0% (= es liegen keine Beziehungen vor) und 100% (= es liegt die maximal mögliche Anzahl Beziehungen vor) annehmen. Die Anzahl maximal möglicher Beziehungen ergibt sich dabei aus der Anzahl Akteure in einem Netzwerk.

  - Cliquenanalyse: Solche Verfahren zielen darauf ab, ein Netzwerk in verschiedene Teilgruppen zu zerlegen. Der Begriff der Clique wird dabei ähnlich verwendet wie in der Umgangssprache: Eine Clique ist eine Gruppe von mindestens drei Personen, die vollständig miteinander verbunden sind.

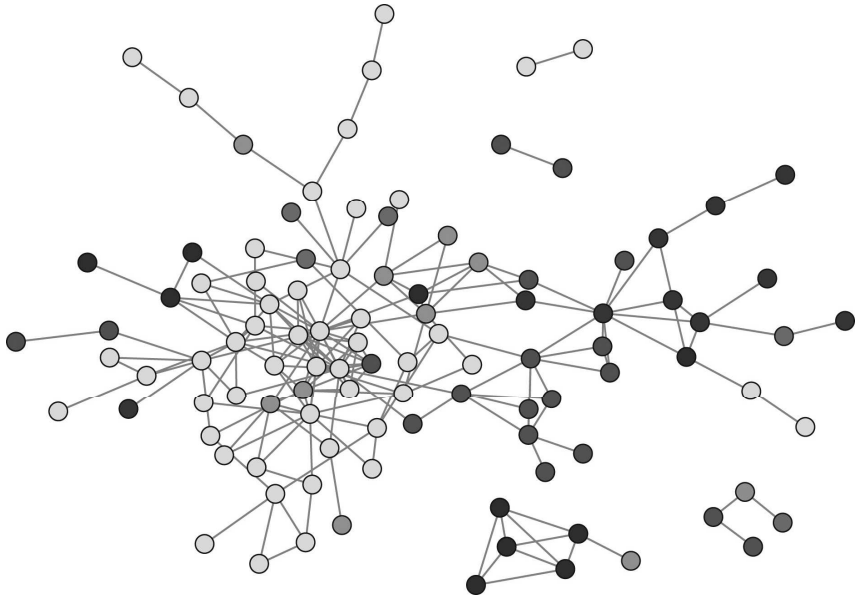
- Akteur: Der Akteur kann in einem Netzwerk ein Unternehmen, ein Individuum, ein Event oder eine soziale Einheit (Personengruppe, Abteilung in einem Unternehmen, Stadt, Nation, etc.) sein.

4. Soziale Netzwerke und Soziale Netzwerkanalyse

4.2. Methoden zur Visualisierung

Multidimensionale Skalierung (MDS):

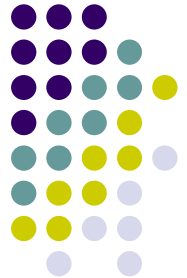
MDS unterteilt man in das intuitiv leichter verständliche „Distance Scaling“ und das mathematisch anspruchsvollere „Classical Scaling“. Bei MDS Verfahren werden die visualisierten euklidischen Distanzen den Pfaddistanzen (wie viele Schritte sind die beiden Knoten im Netzwerk entfernt) angenähert.



- Spring Embedder (Distance Scaling):
  - Algorithmen dieser Gruppe sind iterative Verfahren und haben ihren Namen von der Vorstellung, dass die Kanten eines Graphen durch Federn ersetzt werden. Jede Feder hat eine Federkonstante, also eine optimale Länge. Drückt man die Feder zusammen, drängt die Feder danach, sich wieder auszudehnen. Zieht man die Feder auseinander, will sie sich wieder zusammenziehen.

4. Soziale Netzwerke und Soziale Netzwerkanalyse

4.2. Methoden zur Visualisierung



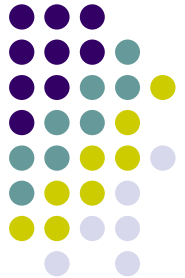
- Spring-Embedder-Algorithmen unterscheiden sich in der Art der implementierten Kräfte. Kamada und Kawai (1989) versuchen die Pfaddistanzen mit den euklidischen Distanzen der Visualisierung anzupassen, indem in jedem Durchlauf der Knoten mit der größten Abweichung in Richtung „bessere“ Position verschoben wird.

- Nachteil ist der Rechenaufwand, sodass Kamada/Kawai auf Netzwerke mit mehreren 1.000 Knoten de facto nicht anwendbar ist.
- Fruchterman und Reingold (1991) haben das Spring-Embedder-Layout-Verfahren insofern optimiert, dass auf jeden Knoten anziehende und abstoßende Kräfte wirken. Angezogen wird ein Knoten von seinen verbundenen Nachbarn, abgestoßen von allen Knoten.

- Ein Vorteil und gleichzeitig Nachteil der Spring-Embedder-Algorithmen stellt nach Eick (1993) die Tatsache dar, dass „Feder-Algorithmen dazu tendieren, Gebiete auszufüllen, da alle Knoten versuchen, sich so nahe zu kommen, wie dies die abstoßenden Kräfte erlauben. Daraus resultiert auch, dass Knoten versuchen, Lücken zu schließen.“ Der Vorteil, der gleichmäßigeren Verteilung der Knoten auf der Fläche mit weniger Überlappungen ist also gleichzeitig der Nachteil der Verfälschung der Struktur.

4. Soziale Netzwerke und Soziale Netzwerkanalyse

4.2. Methoden zur Visualisierung



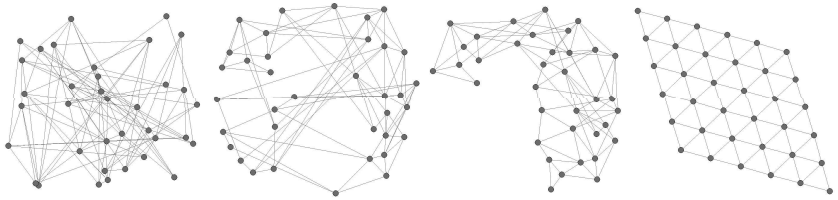
- Multidimensional Scaling (Classical Scaling):
  - Verfahren der klassischen multidimensionalen Skalierung (MDS) sind statistische Verfahren zur Messung von Ähnlichkeiten und Unähnlichkeiten und kommen aus dem Bereich der Multivariaten Statistik. Bei der Visualisierung von sozialen Netzwerken wird die Matrix der Ähnlichkeiten bzw. Unähnlichkeiten, welche die Ausgangsbasis für MDS-Verfahren darstellt (z.B. durch die Berechnung der Pfaddistanzen) ermittelt.
  - Ziel: Verfahren ist eine niedrigdimensionale (zwei oder drei) Visualisierung höherdimensionaler Daten
  - Im Gegensatz zu den Distance Scaling Ansätzen haben klassische MDS Verfahren ein globales Optimum, das direkt errechnet wird. Das heißt, dass die Berechnung für ein Ausgangsnetzwerk immer zum gleichen Ergebnis führt.
  - Nachteil: Auch MDS-Verfahren sind durch quadratischen Aufwand in ihrer Skalierbarkeit beschränkt.

4. Soziale Netzwerke und Soziale Netzwerkanalyse

4.2. Methoden zur Visualisierung

- Factor Analysis - Singular Value Decomposition:
- Auch hierbei handelt es sich um Verfahren, die aus der multivariaten Statistik kommen.
- SVD transformiert N Variablen in n neue Variablen oder Dimensionen. Die unterschiedlichen SVD Ansätze werden immer gleich berechnet, unterscheiden sich nur in der Art, wie die Datenmatrix vorbereitet wird.

- Kriterium:
- Korrelation der Pfaddistanzmatrix mit der Matrix der euklidischen Distanzen
  - Das heißt, dass die Distanzen in der Visualisierung zwischen allen Knotenpaaren mit den Pfaddistanzen verglichen werden. Sind jene Knoten, die direkt miteinander verbunden sind, auch in der Visualisierung tatsächlich nebeneinander? Der so erhaltene Korrelationskoeffizient r^2 gibt im Bereich von -1 bis +1 die „Richtigkeit“ der Visualisierung an. Je höher das r^2 desto besser spiegelt das Layout die tatsächliche Struktur wider.



5. Diskussion