REFERENCES

ARABIE, Phipps, Scott A. BOORMAN und Paul R. LEVITT, 1978: Constructing blockmodels: How and why. In: Journal of Mathematical Psychology 17: 21 – 63.

BREIGER, Ronald L., Scott A. BOORMAN und Phipps ARABIE, 1975: An algorithm for clustering relational data, with application to social network analaysis and comparison with multidimensional scaling. In: Journal of Mathematical Psychology 12: 328 – 383.

COLSON, Elisabeth, 1963: The Plateau Tonga of Northern Rhodesia. Manchester. Manchester University Press for Rhodes – Livingstone Institute.

LIGHT, John M. and Nicholas C. MULLINS, 1979: A primer on blockmodelling procedure. In: Holland, P.W. und S. Leinhardt (eds.): Perspectives on social network research. London, Academic Press:85 – 118.

Codierung und Gruppierung von Netzwerkdaten

ELMAR KLEMM und RAFAEL WITTEK

EINLEITUNG

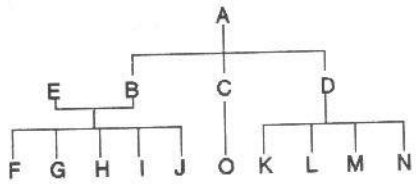
Netzwerkdaten sind Beziehungsdaten. Gegenstand des vorliegenden Beitrages ist die Aufbereitung – Codierung – und statistische Auswertung von Beziehungsdaten. Dabei soll zwei Punkten besondere Beachtung geschenkt werden. Zunächst wird aufgezeigt, welche Möglichkeiten bestehen, Beziehungsdaten statistisch handhabbar zu machen. Dabei werden aus der Vielzahl von Möglichkeiten drei ausgewählt und deren Unterschiede herausgearbeitet. Im zweiten Teil beschäftigen wir uns mit der Clusteranalyse. Durch dieses statistische Gruppierungsverfahren können komplexe Beziehungsdaten übersichtlich aufbereitet werden. Hierbei legen wir besonderen Wert auf die Herausarbeitung des Zusammenhangs zwischen der Codierung und den Ergebnissen der Clusteranalyse. Desweiteren wollen wir in diesem Teil nicht nur ein uniplexes Netzwerk analysieren, das eine Beziehung zum Gegenstand hat, sondern auch aufzeigen, auf welche Art ein multiplexes Netzwerk, welches durch mehrere Beziehungen gekennzeichnet ist, bearbeitet werden kann. Die gewonnenen allgemeinen Erkenntnisse werden exemplarisch angewandt. Als Beispiel dient uns ein Netzwerk, welches wir einer stadtethnologischen Untersuchung Roger Sanjeks (1982:78) entnommen haben (Abbildung 1). Es besteht aus 69 Personen aus dem Wohngebiet Adabraka in Accra, der Hauptstadt Ghanas. Für diese Akteure sind von Sanjek drei verschiedene Arten von Beziehungen festgehalten worden: Verwandtschaftliche Beziehungen, Koch – und Eβbeziehungen (ob die Personen einer Kochgemeinschaft angehören) und Wohnbeziehungen (ob die Personen im selben Haushalt leben).

DIE AKTEURE UND IHRE BEZIEHUNGEN: CODIERUNGSPROBLEME

Unser Netzwerk wird als Person x Person Matrix (lies: Person mal Person Matrix) codiert. Da Beziehungen Gegenstand unserer Untersuchung sind, stehen in den einzelnen Feldern der Matrix Werte, die die Beziehung der Person in der Spalte mit der Person in der Zeile charakterisieren. Welche Werte für die zu erfassenden Beziehungen ver-

Für die formale Analyse, die wir hier vornehmen wollen, verfügen wir nicht über die emischen Kategorien, sondern lediglich über die in Form eines Verwandtschaftsdiagramms vom Ethnographen festgehaltene Information. Diese gilt es so genau wie möglich zu codieren. Um dies zu erreichen gehen wir folgendermaßen vor: Jede Person, die in dem Verwandtschaftsdiagramm über eine Linie direkt mit einer anderen Person verbunden ist, bekommt mit dieser anderen Person eine „1" und ansonsten eine „0".

Abbildung 2: Schema eines Verwandtschaftsnetzes



In Abbildung 2 würde die Beziehung zwischen A und B (A,B) mit einer „1" codiert werden. Auch das Verhältnis (B,D) würde als „1" wiedergegeben werden. Nicht jedoch das Verhältnis (A,H), das hier eine „0" erhalten würde, da diese beiden Personen nicht direkt durch eine Linie verbunden sind.

Ein Punkt verdient bei der binären Codierung besondere Betonung, die Behandlung der Diagonalfelder. Eine Beziehungsmatrix enthält nicht nur Felder, in denen der Wert für die Beziehung zwischen zwei Personen festgelegt ist, sondern unvermeidbar auch Felder, in denen die Werte der Beziehungen der Personen zu sich selbst enthalten sind. Diese Felder kann man nicht nach dem oben beschriebenen Schema codieren, da Beziehungen zwischenmenschlich sind. Normalerweise weist man diesen Feldern den Wert „0" zu. Es gibt jedoch statistische Verfahren, die bessere Ergebnisse erzielen, wenn man diese Felder mit einer „1" besetzt. Im Abschnitt „Gruppierung eines multiplexen Netzwerks" werden wir einen solchen Fall vorstellen.

Wie wir in diesem Abschnitt gesehen haben, erfaßt die binäre Codierung das direkte Beziehungsumfeld einer jeden Person. Im Gegensatz dazu wird bei der geodesischen Codierung eher die Stellung einer Person in Relation zum ganzen Netzwerk festgehalten.

Geodesische Codierung

Bei der geodesischen Codierung werden die kürzesten Pfadlängen einer Person zu allen anderen Personen im Netzwerk codiert. Unter Pfadlängen versteht man eine Angabe über die Zahl der Beziehungen, die

zwischen zwei Personen liegen. In einer normalen geodesischen Matrix befindet sich in jedem Feld ein Wert, der angibt, wieviele Beziehungen im kürzesten Fall zwischen der Person in der Zeile und der Person in der Spalte liegen. In Abbildung 2 würde die Beziehung (J,N) den Wert „3" bekommen, da zwischen J und N die Beziehungen (J,B), (B,D) und (D,N) liegen. Der Wert für (J,B) wäre „1" würde also dem Wert der binären Codierung entsprechen. Die binäre Codierung kann man als einen Spezialfall der geodesischen Codierung betrachten, bei der nur die Pfadlängen von „1" in die Matrix aufgenommen werden. Zwei Punkte sollen hier hervorgehoben werden:

(1) Ein Problem bei der Bestimmung der Pfadlängen ist die Vergabe eines Wertes für Personen, die unerreichbar sind. Eigentlich sollten diese Personen den Wert „unendlich" erhalten, dies ist jedoch praktisch kaum zu vertreten. Das Vorhandensein einer Verbindung würde im Gegensatz zum Fehlen einer Verbindung zur rechnerischen Bedeutungslosigkeit herabsinken. Im Falle des Verwandtschaftsnetzwerkes aus Adabraka betrug die maximale geodesische Pfadlänge „8". Gut interpretierbare statistische Rechenergebnisse (d.h. Clusteranalysegruppierungen, die weder die Feinstruktur von Clustern verwischten noch unerklärliche Zuordnungen von Clustern vornahmen) erhielten wir, wenn wir den unerreichbaren Personen den Wert „10" zuwiesen.

(2) Die Diagonalfelder in geodesischen Matrizen werden mit dem Wert „0" besetzt.

Das Konzept der geodesischen Codierung ähnelt Verfahren, welche die Verwandtschaftsethnologie zur Bestimmung genealogischer Distanzen entwickelt hat (Ballonoff 1976:10). Inwiefern diese explizit verwandtschaftsethnologischen Distanzmaße, wie etwa das römische oder das kanonische, geeignet wären, an dieser Stelle Verwendung zu finden, kann hier jedoch nicht geklärt werden.

Wahrscheinlichkeitscodierung

Die letzte Art der Codierung, die hier vorgestellt werden soll, ist die Wahrscheinlichkeitscodierung. Die Wahrscheinlichkeitscodierung besitzt Ähnlichkeit zur binären Codierung, berücksichtigt aber bei jeder Beziehung, die codiert wird, die Tatsache, daß die daran beteiligten Personen auch Beziehungen zu anderen Personen haben. Dieser Codierung liegt also die Annahme zugrunde, daß die Anzahl der Beziehungen, die eine Person unterhält, Einfluß auf die Beziehungen hat. Besteht ein Netzwerk wie in Abbildung 2 aus 15 Personen und es ist bekannt, daß eine Person genau eine Beziehung hat, so ist die Wahrscheinlichkeit, daß diese Person eine Beziehung zu einer bestimmten anderen Person (= 14 potentielle Personen) aus dem Netz hat 1/14 (Person O im Beispielnetzwerk). Unter der Annahme, daß die Ereignisse voneinander unab-

schaftsmatrix kann die Begründung lauten: Personen, die sich verwandtschaftlich nahe stehen, sind sich ähnlich. Die geodesische und die probabilistische Matrix bilden die Grundlage für die im nächsten Kapitel folgende Analyse des Verwandtschaftsnetzes.

Das Rechenverfahren

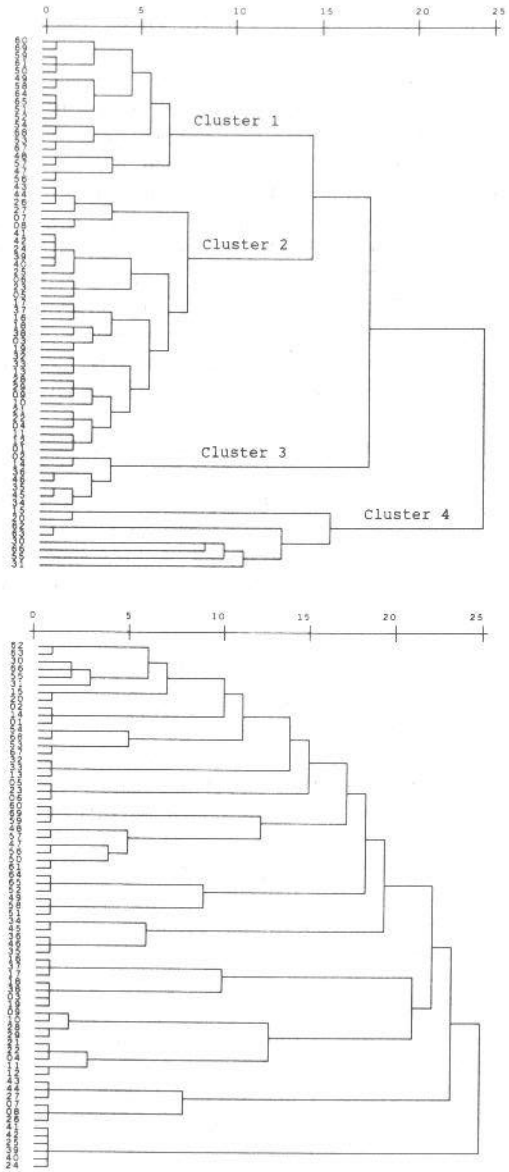
Die von uns verwendeten Clusteralgorithmen sind das „Ward – Verfahren" sowie „Average Linkage Between Groups". Beide gehören zur Gruppe der hierarchischen Clusteranalysen. Ziel einer hierarchischen Clusteranalyse ist es, Cluster sukzessive durch Fusion der ähnlichsten Objekte oder Personen aufzubauen. Auf der untersten Ebene, der feinsten Partitionierungsstufe, bildet jede Person in der Matrix einen Cluster für sich. Nach bestimmten Kriterien werden iterativ, d.h. in sich wiederholenden Rechenschritten, jeweils die ähnlichsten Cluster miteinander verbunden. Bei der Ward – Methode werden jene Cluster verschmolzen, bei deren Fusion sich die geringste Erhöhung der Fehlerquadratsumme ergibt; „Average Linkage between Groups" fusioniert jene Cluster, die die größte durchschnittliche Ähnlichkeit besitzen (auf die Details der Clusteranalyse kann hier aus Platzgründen nicht eingegangen werden; vgl. dazu Bortz 1985:697 – 702, Eckes/Roßbach 1980). Durch die Zuordnung von Personen zu Clustern wird die Interpretation komplexer Beziehungsdaten erleichtert. Auf diese Weise können auch Informationen, die „mit bloßem Auge" gar nicht oder nur schwer erkennbar sind, offengelegt werden. Dies soll nun zunächst anhand des vorliegenden Verwandtschaftsnetzes aus unserem Fallbeispiel exemplarisch dargestellt werden.

GRUPPIERUNG EINES UNIPLEXEN NETZES: DAS VERWANDTSCHAFTSNETZWERK IN ADABRAKA

Grundlage der Clusteranalyse des Verwandtschaftsnetzes sollen zwei Ursprungsmatrizen bilden, die geodesische und die probabilistische. Hinter den beiden Distanzmatrizen stehen unterschiedliche Konzepte, das Verwandtschaftsdiagramm abzubilden. Die Frage ist nun, wie sich diese beiden Konzepte jeweils auf die Gruppierung des Netzes auswirken. Wenn beide Matrizen demselben Rechenverfahren unterzogen werden und die Ergebnisse voneinander abweichen, so sind diese Unterschiede auf die jeweilige Matrix zurückzuführen. Bereits ein Blick auf die Gesamtstruktur der beiden Dendrogramme läßt erkennen, daß die Ergebnisse voneinander abweichen (siehe Abbildung 3).

Die auf der Wahrscheinlichkeitscodierung beruhende Gruppierung fügt die auf den niedrigen Fusionsstufen identifizierten Teilcluster sukzessive aneinander. Die auf unteren Ebenen gefunden Cluster werden also meist

Abbildung 3: Dendrogramme der Clusteranalysen des Verwandtschaftsnetzwerkes: mit geodesischer Matrix (oben), mit Wahrscheinlichkeitsmatrix (unten).



steranalyse die Gruppierung vor allem von den strategisch wichtigen Personen des Netzwerkes determiniert wird; bildet die Wahrscheinlich – keitsmatrix die Grundlage, werden Beziehungstypen erkannt.

GRUPPIERUNG EINES MULTIPLEXEN NETZES: VERWANDTSCHAFT, KONSUM UND RESIDENZ IN ADABRAKA

Im vorangehenden Teil wurde jeweils nur eine Matrix, die lediglich einen Beziehungstyp codierte, betrachtet. Will man mehrere Bezie – hungstypen zusammen analysieren, muß man sich eines leicht veränder – ten Vorgehens bedienen. In unserem Fall wollen wir drei verschiedene Beziehungstypen und damit drei verschiedene Netzwerke auf einmal analysieren. Es handelt sich dabei erstens um das schon behandelte Verwandtschaftsnetzwerk (wer ist mit wem verwandt) zweitens um ein Netzwerk, das die Koch – und Eßbeziehungen erfaßt (wer kocht mit wem) und drittens um ein Netzwerk, das die Wohnbeziehungen (wer wohnt mit wem in derselben Wohnung) darstellt. Vereinfacht ausge – drückt, werden die drei zu erfassenden Matrizen zu einer Matrix kom – primiert, indem man aus den Zeilen der Matrizen eine Ähnlichkeitsma – trix erzeugt. Die so erzeugte neue Matrix wird dann einem bestimmten Gruppierungsverfahren unterworfen. Bei der Konstruktion der neuen Matrix stehen dabei mehrere Möglichkeiten zur Verfügung. Gegenstand dieses Kapitels wird es sein, aufzuzeigen, wie sich diese verschiedenen Möglichkeiten auf die Gruppierung auswirken, indem sie bestimmte strukturelle Aspekte eines Netzwerkes unterschiedlich hervorheben.

Die Bildung von Ähnlichkeitsmatrizen

Als Ausgangspunkt für die nun folgende Analyse verwenden wir die binäre Matrix. Die geodesische Matrix und die Wahrscheinlichkeitsmatrix bieten in diesem Fall gegenüber der binären Matrix keine wesentlichen Erkenntnisvorteile. Im Falle der geodesischen Codierung deshalb, weil die geodesischen Werte in der Koch – und der Wohnmatrix entweder auf „1" oder auf „unerreichbar" stehen. Dies ist dadurch bedingt, daß erstens die Personen im Koch – und Wohnnetzwerk jeweils eindeutig einer Gruppe zugeordnet sind (z.B. dem Haushalt in dem sie wohnen) und die Gruppen untereinander unverbunden sind, und zweitens die Gruppen selbst intern maximal verbunden sind. In einem so gearteten Netzwerk erhalten alle Personen einer Gruppe gegenseitig eine „1" und alle Personen verschiedener Gruppen erhalten gegenseitig den Wert für „unerreichbar". Insofern kann die Stärke einer geodesischen Matrix, den Beziehungsfernbereich einer jeden Person differenziert zu erfassen, nicht zum Tragen kommen. Ähnliches läßt sich für die Wahrscheinlichkeits – codierung sagen. Da die Gruppierungen der Koch – und Wohnnetzwerke

immer maximal intern verbunden sind, sind die Werte, welche die einzelnen Personen einer Gruppe zueinander erhalten, alle identisch. Besteht eine Gruppe z.B. aus zehn Personen erhalten alle Personen für die Beziehungen zueinander den Wert 1/81 (= 1/9 * 1/9). D.h. die Stärke der Wahrscheinlichkeitscodierung, den Beziehungsnahbereich jeder Person zu differenzieren, entfällt, da alle Personen, die zum Nahbereich einer Person gehören, denselben Wert erhalten. Da diese beiden Codierungen gegenüber der binären Codierung aufgrund des parzellierten Aufbaus der Beziehungsnetze und der Beziehungshomogenität der isolierten Subgruppen keinen Erkenntnisgewinn bringen, benutzen wir bei der nun folgenden Analyse die herkömmliche Form der Codierung, die binäre Codierung.

Um die Informationen, die uns alle drei Netzwerke bieten, auf einmal verrechnen zu können, müssen wir die drei einzelnen Matrizen zu einer einzigen vereinen. Zu diesem Zweck legen wir die drei Matrizen (3 x n x n) einfach hintereinander und erzeugen dadurch eine Matrix (n x 3n). Diese Matrix besteht jetzt aus 69 Zeilen und 207 Spalten (3 x 69). Da die Clusteranalyse eine quadratische Matrix als Eingabe erwartet (Zahl der Zeilen ist gleich der Zahl der Spalten), in der die Information über die Ähnlichkeit jeder Person mit jeder Person codiert ist, muß die jetzt vorliegende rechteckige Matrix (n x 3n) in eine qua – dratische Ähnlichkeitsmatrix umgewandelt werden. Diese Ähnlichkeits – matrix wird aus den Zeilen der aggregierten Matrix gebildet, so daß eine 69 x 69 Matrix entsteht. Zur Konstruktion einer solchen Ähnlich – keitsmatrix stehen eine ganze Reihe von Maßen zur Verfügung, deren Verwendung jeweils spezifische Skalenniveaus voraussetzt (Bortz 1985:28 – 33). So gibt es auch für binäre Daten (0 – 1 – Daten) eigens konstruierte Maße. Wie binäre Ähnlichkeitsmaße allgemein errechnet werden soll ein Beispiel erläutern. Die folgende binäre Matrix enthält die Informationen, ob die Personen A und B eine Verwandtschaftsbe – ziehung zu den Personen A bis O besitzen oder nicht:

Tabelle 1: Ausschnitt einer binären Verwandtschaftsmatrix

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
B	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0

Wie man sieht, finden sich in der Matrix vier mögliche Kombinatio – nen von Einsern und Nullen:

a = (1,1) – A und B sind mit der entsprechenden Person verwandt,
b = (1,0) – A ist mit ihr verwandt und B nicht,
c = (0,1) – B ist mit ihr verwandt und A nicht,

werden konnten. Diese Ähnlichkeitsmatrizen wurden dann in SPSS ein-gelesen und einer Clusteranalyse (Average Linkage Between Groups) unterzogen. Die dabei auftretenden unterschiedlichen Gruppierungser-gebnisse müssen Effekte der beiden unterschiedlichen Ähnlichkeits-koeffizienten sein, da die Gruppierungsrechnung in beiden Fällen gleich ist. Wieso die beiden unterschiedlichen Ähnlichkeitskoeffizienten zu verschiedenen Ergebnissen führen, soll im folgenden Abschnitt geklärt werden.

Auswirkungen unterschiedlicher Ähnlichkeitsmatrizen

Zwei aus unterschiedlichen Ähnlichkeitskoeffizienten berechnete Ähn-lichkeitsmatrizen wurden demselben Clusteranalyseverfahren unterzogen. Dabei weichen die Ergebnisse voneinander ab. Ihre Interpretation wird um ein vielfaches erleichtert, wenn bekannt ist, in welcher Weise sie sich voneinander unterscheiden. Unterschiede in der Gruppierung, die durch unterschiedliche Ähnlichkeitskoeffizienten bedingt sind, müssen dadurch begründet sein, daß ein und dasselbe Personenpaar für seine Ähnlichkeit vom Similarity S einen anderen Wert zugewiesen bekommt, als von Driver-Kroebers G. Eingehende Analysen dieses Verhältnisses, auf die hier nicht näher eingegangen werden soll, führten zu dem Er-gebnis, daß das Verhältnis von G zu S von zwei Faktoren abhängt:

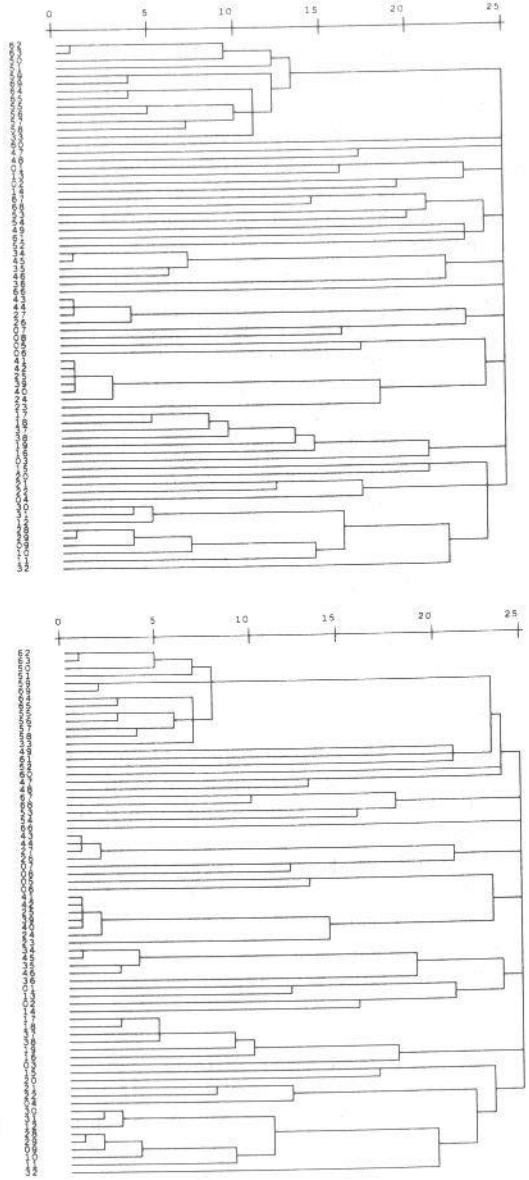
(1) Von der Anzahl der „1-1"-Kombinationen. Je höher die Anzahl der „1-1"-Kombinationen wird, desto eher stimmen die Werte von G und S exakt überein.

(2) Vom Verhältnis der Beziehungsanzahl der ersten Person zur Be-ziehungsanzahl der zweiten Person. Je krasser das Verhältnis zwischen den Beziehungsanzahlen der beiden Personen ist, umso größer wird G im Vergleich zu S.

Dies hat natürlich für eine Clusteranalyse weitreichende Konsequen-zen. Die oben genannten Einflußfaktoren wirken sich auf Personenver-gleiche (die Errechnung der Ähnlichkeit zwischen zwei Personen) dann aus, wenn die Personen folgende Bedingungen erfüllen: (1) Die beiden Personen müssen eine relativ geringe Anzahl von „1-1"-Kombinatio-nen haben, dabei darf die Anzahl der „1-1"-Kombinationen jedoch nicht gleich Null sein. D.h. beide Personen müssen im Beziehungsnetz entweder miteinander eine direkte Beziehung unterhalten oder höchstens eine Pfadlänge von „2" voneinander entfernt sein. Die Personen müssen also relativ nahe im Beziehungsnetz zusammen liegen. (2) Beide Perso-nen müssen über eine stark unterschiedliche Anzahl von Beziehungen verfügen. Anders ausgedrückt: Eine Person muß eine relative zentrale Stellung im Beziehungsnetz einnehmen, wohingegen die andere Person relativ peripher bezüglich des Beziehungsnetzes sein muß.

Sind diese beiden Bedingungen erfüllt, ist Driver-Kroebers G dieser beiden Personen eindeutig höher als das Ähnlichkeitsmaß Similarity S.

Abbildung 4: Dendrogramme der Clusteranalysen aller drei Beziehungs-netzwerke: mit Similiarity S (oben), mit Driver-Kroebers G (unten).



Netzwerkanalyse

Ethnologische Perspektiven

Herausgegeben von
Thomas Schweizer

DIETRICH REIMER VERLAG
BERLIN