

# L'ANALYSE FACTORIELLE DANS L'ÉLABORATION DES ENTRÉES DE DICTIONNAIRE

UGO BERNI CANANI

Je vous ai parlé précédemment d'une curiosité liée à la série de Waring; je vous en présenterai maintenant une autre liée à l'analyse factorielle. Il s'agit d'une application de cette technique en vue de la rédaction d'articles de dictionnaire.

Supposons que nous ayons un ensemble de mots associés d'une façon paradigmatique au mot devant faire l'objet de l'article. On peut les représenter comme les noeuds d'un graphe en reliant entre eux ceux entre lesquels il y a une relation paradigmatique. S'il y a beaucoup de mots et beaucoup d'associations le résultat sera un graphe assez compliqué. Le problème est de le dessiner de la façon la plus claire possible, en essayant de minimiser les enchevêtrements, de façon à mettre en évidence les différentes aires sémantiques qui peuvent exister dans l'ensemble. Ce n'est pas simple. Il y a même un article d'un illustre spécialiste d'analyse combinatoire, W.T. Tutte, intitulé *Comment dessiner un graphe*. Or l'analyse factorielle se révèle justement un instrument très utile pour donner une bonne représentation des graphes. Voici quelques exemples. Le premier est constitué par un ensemble de termes associés à la notion de droit. Le graphe correspondant a été traité avec deux méthodes d'analyse factorielle: l'analyse des correspondances et l'analyse en composantes principales. Sur la planche 1 on peut voir le résultat obtenu

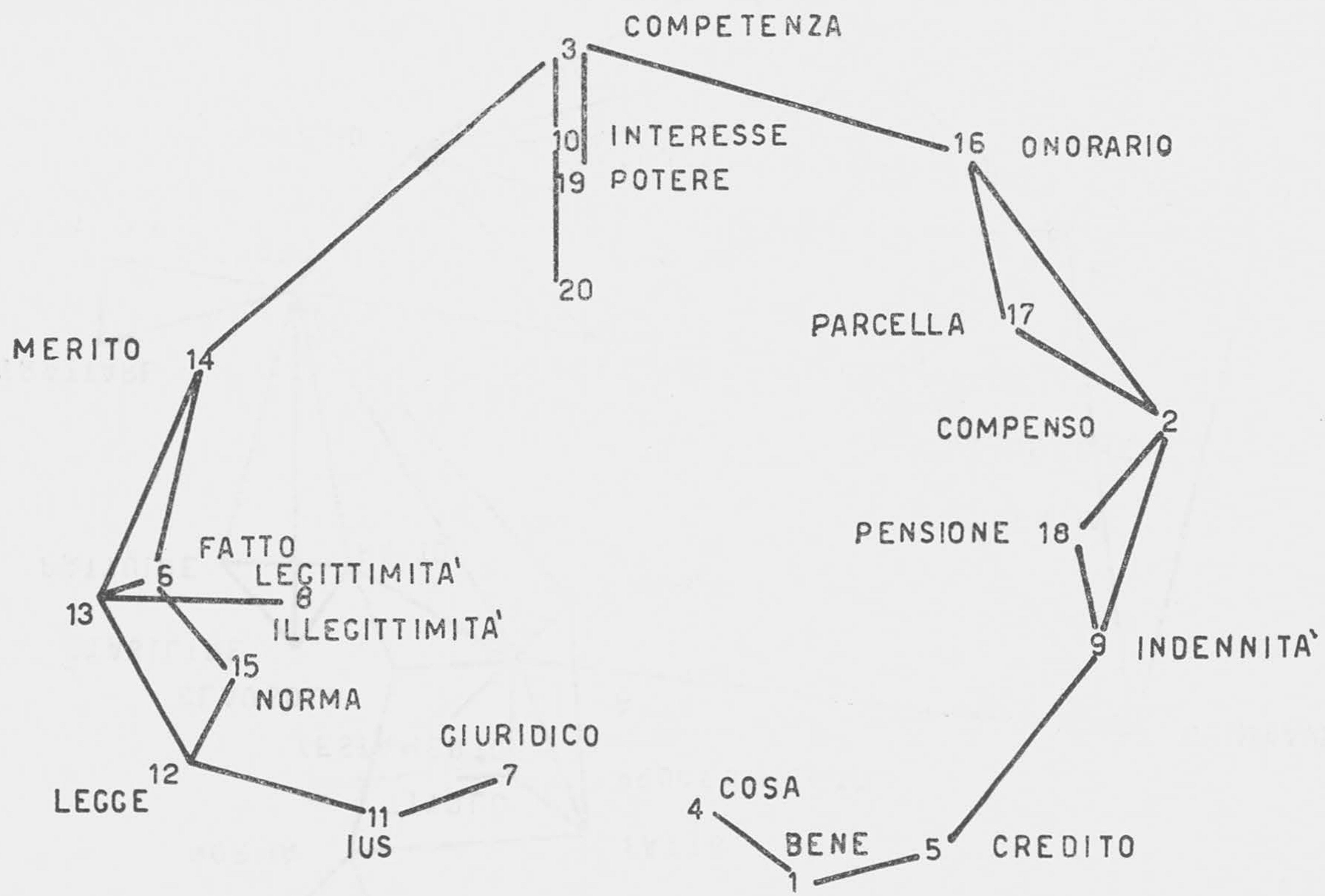
avec l'analyse des correspondances sur le plan déterminé par les deux premiers axes factoriels. On distingue très bien sur le côté droit une région formée par les termes *giuridico*, *jus*, *legge*, *norma*, *legittimità*, *illegittimità*, puis *fatto* et *merito*; en haut une autre où apparaissent les mots *competenza*, *interesse*, *potere*; sur la gauche les mots *onorario*, *parcella*, *compenso*; enfin des termes comme *credito*, *bene*, *cosa*. On a un dessin qui se présente souvent dans ce type d'analyse et qu'on pourrait appeler une méduse vertébrée. En effet, d'une masse, d'un squelette central, partent plusieurs tentacules. La deuxième planche montre une analyse en composantes principales du même graphe. J'y vois encore sur le premier plan factoriel une méduse, mais cette fois une méduse crustacée: les noeuds les plus denses forment un cercle et les tentacules sont dirigées vers l'intérieur. On peut y retrouver les mêmes régions que dans la première planche. Le graphe qu'on a vu était peut-être trop simple. Passons à un autre plus compliqué. Il s'agit d'un ensemble de mots associés au mot *contratto*. J'ai réussi, après pas mal d'efforts, à dessiner le graphe que vous voyez sur la planche 3. J'ai même sauté quelques mots pour ne pas le rendre illisible. On peut y voir quand même une région formée par des termes tels que *negozio*, *negoziale*, *contrattuale*, *preliminare*; une autre où le sens de *contratto* tend vers les notions de *atto* à travers *accordo*, *patto*, *rapporto*; encore une autre région où sont représentés des contrats spécifiques etc. Je voulais séparer ces différentes régions et j'ai essayé de traduire ce que j'aurais fait de façon intuitive. Je me suis dit: « dans ce dessin je peux voir une sorte de saucisson dont on a relié les deux extrémités par erreur »; un saucisson n'est peut-être pas l'image la meilleure, peut-être vaudrait-il mieux parler d'un rôti de veau, d'un rosbif mal ficelé. Il y a en effet une ficelle qui relie deux points qui, dans le morceau, sont à l'opposé: *atto* et *compravendita*. Je voudrais mettre en évidence cette ficelle qui me semble le premier obstacle pour une bonne

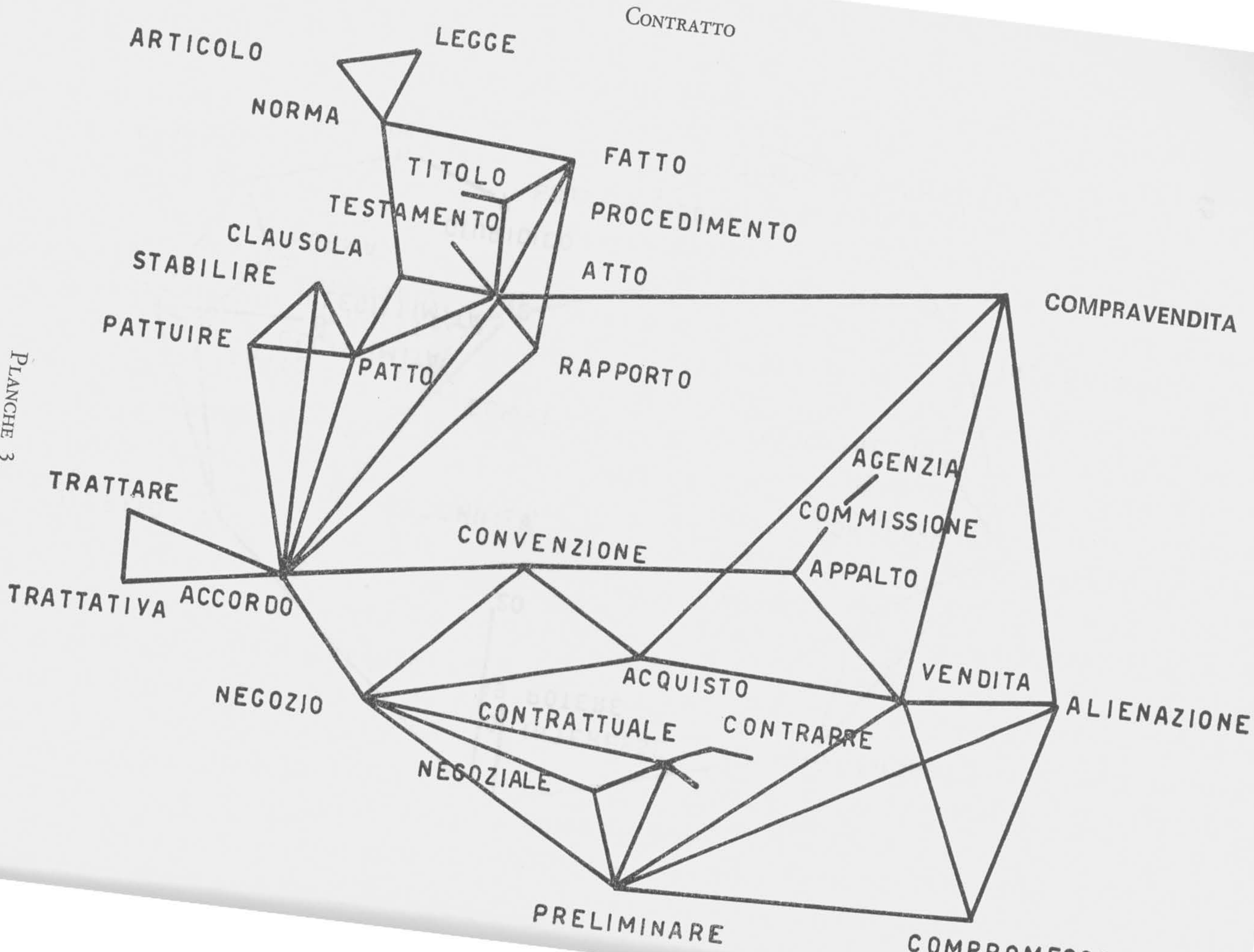
séparation des régions. Le premier essai (planche 4) est fait avec l'analyse des correspondances. On retrouve la méduse vertébrée. Les mots sont représentés par des chiffres: ceux qui sont le plus fortement reliés entre eux se trouvent au centre d'où partent quelques filaments vers des mots périphériques (*testamento, commissione, agenzia* etc.). *Atto* et *compravendita* sont, hélas, dans la région centrale représentés respectivement par les numéros 7 et 11. Pour les mettre en évidence je devrais couper les filaments périphériques et répéter l'analyse sur la seule masse centrale. Sur la planche 5 on voit le résultat de l'analyse en composantes principales. On a, comme on pouvait s'y attendre, une méduse crustacée où tous les points importants sont sur la surface extérieure, les tentacules tendant vers l'intérieur. Le chemin le plus long est exactement celui que je voulais couper: il va du noeud 7 au noeud 11. Si on coupe ce lien et qu'on réitère ce procédé on arrivera à simuler automatiquement la répartition que j'aurais faite intuitivement du rosbif. Avec un procédé un peu plus compliqué on peut forcer la situation. Si on applique l'analyse à une matrice dont les lignes représentent les noeuds du graphe et les colonnes les composantes complètes maximales auxquelles ils appartiennent, on a avec l'analyse en composantes principales le résultat figurant sur la planche 6. Le chemin qui relie le point 7 au point 11 est nettement séparé de tous les autres.

Sans doute s'agit-il de 'bricolage' factoriel. Mais il me semble que l'utilisation de procédures automatiques de ce genre pourrait faciliter la tâche de celui qui avec l'analyse de groupes de mots ou de contextes doit rédiger les articles d'un dictionnaire.



PLANCHE 2.





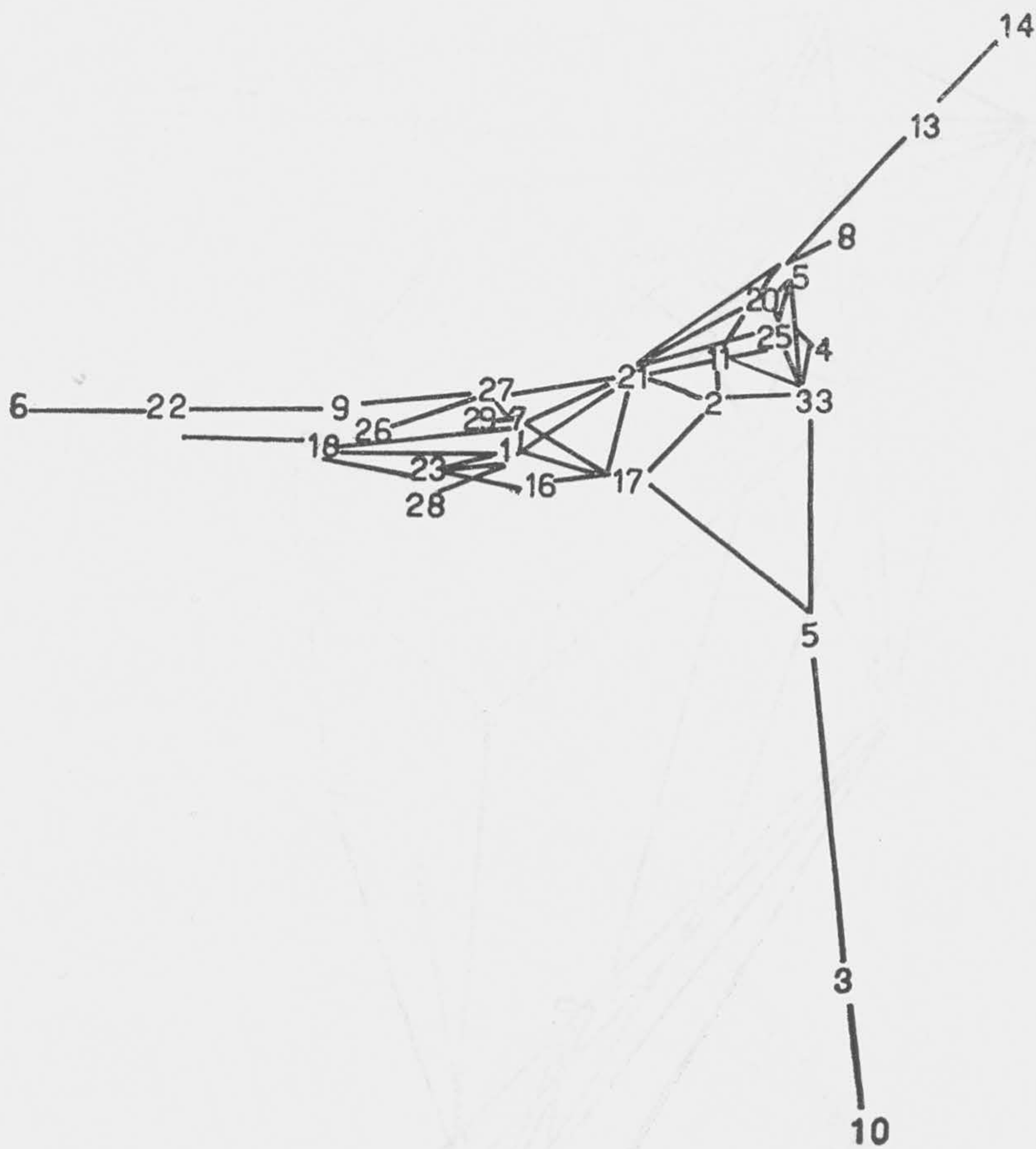


PLANCHE 4.

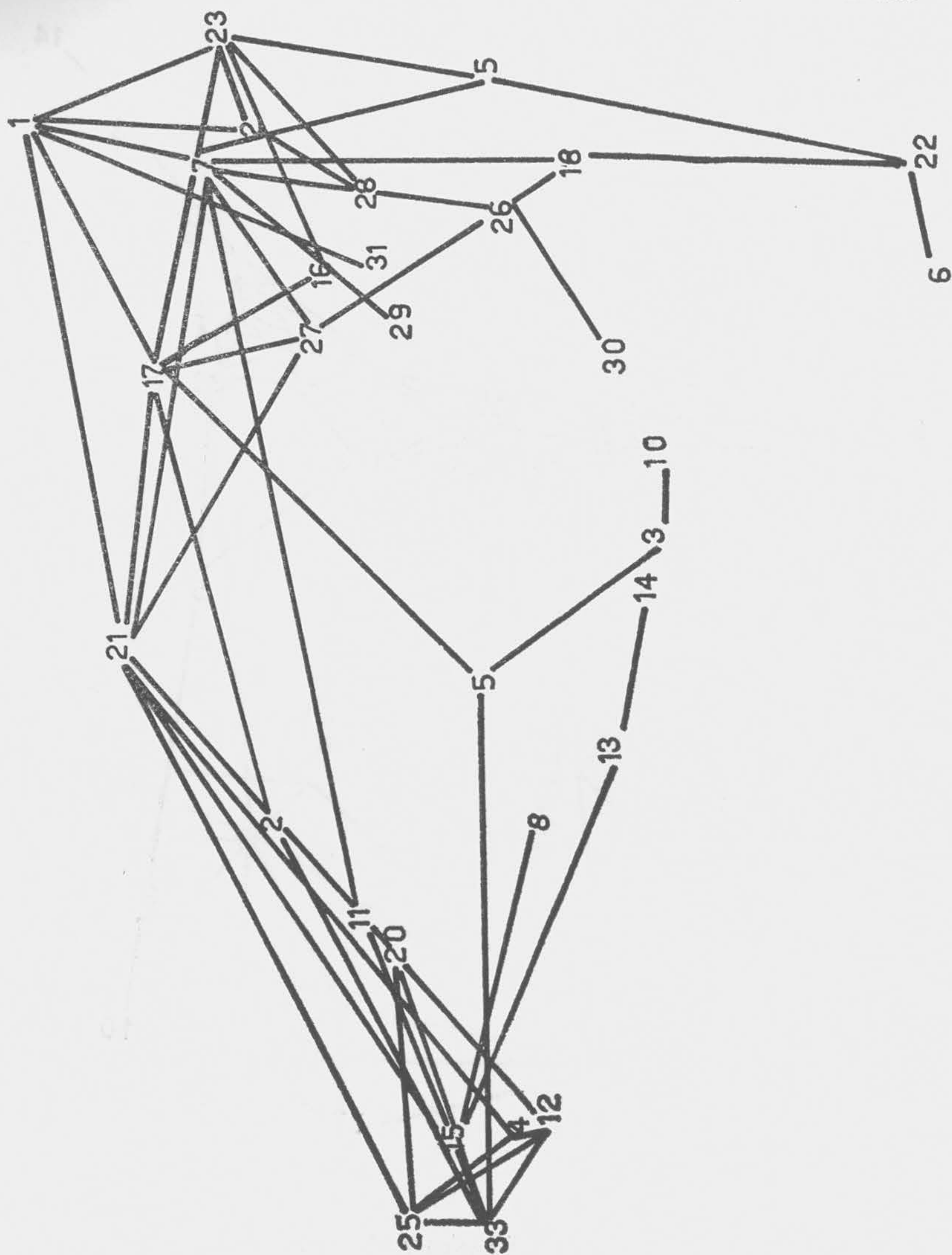


PLANCHE 5.

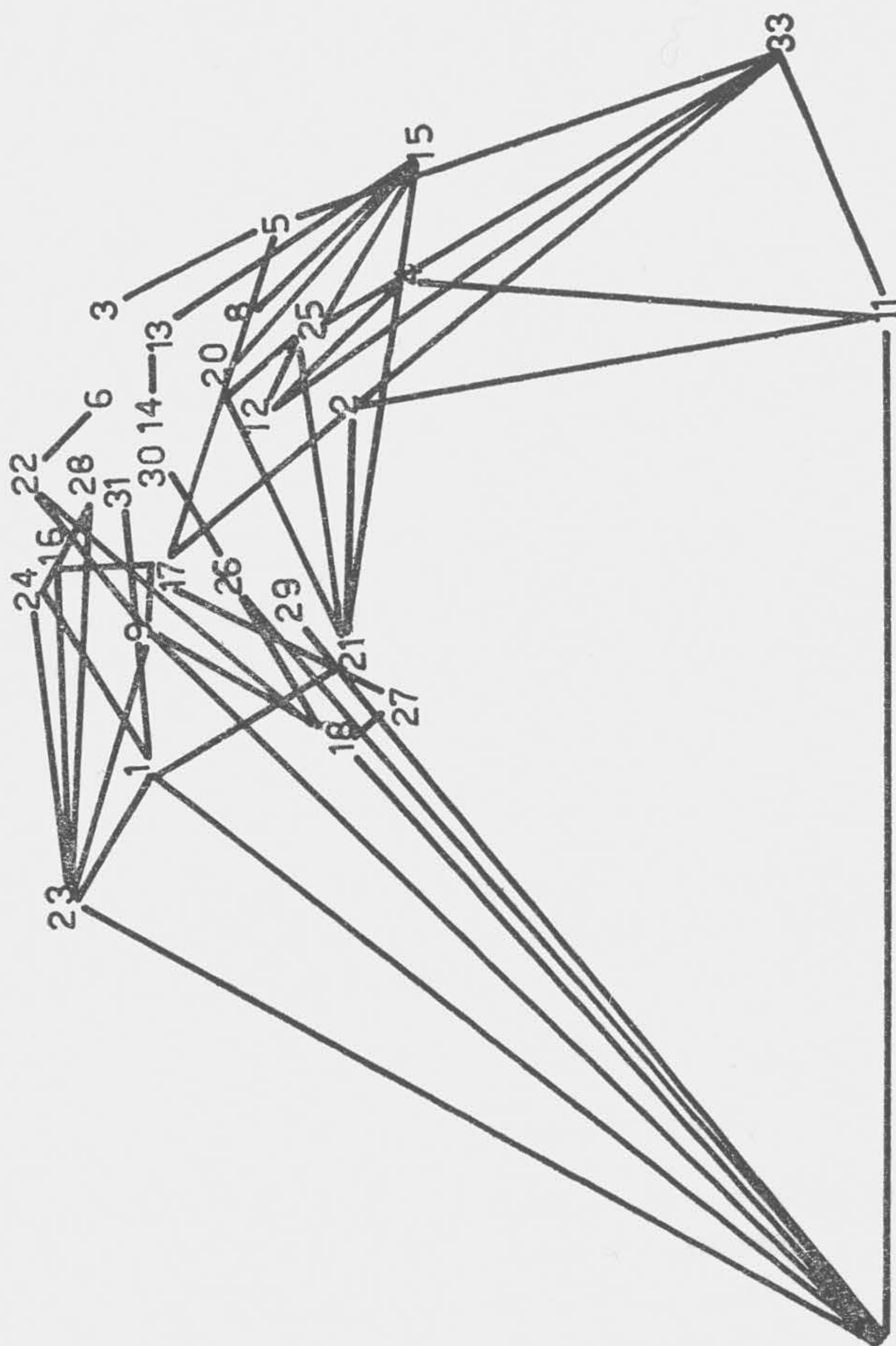


PLANCHE 6.