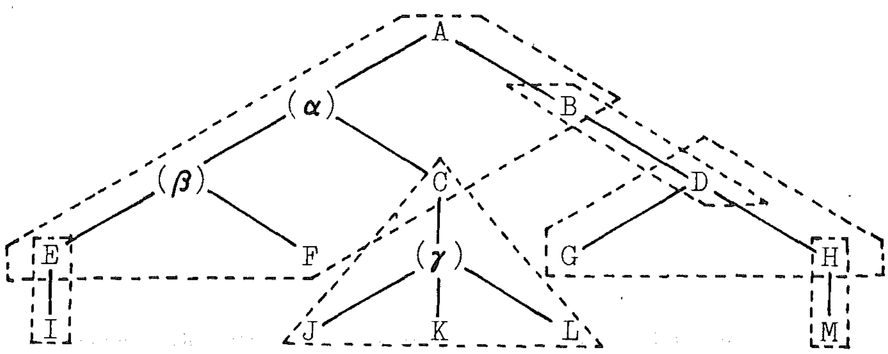
A comparison of local genealogical algorithms and conventional methods of textual criticism

by

D. NAJOCK

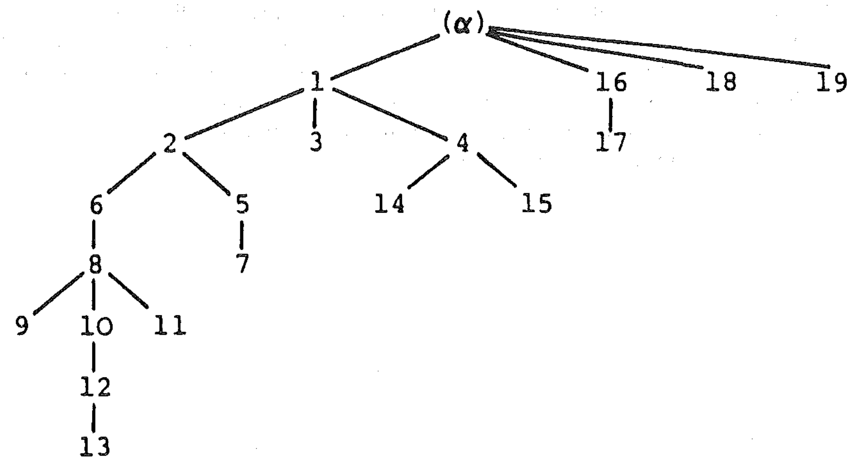
Freie Universität Berlin - DEUTSCHLAND

Conventional textual criticism is not so far from local genealogical algorithms as it might seem at first sight. A rough type of localizing has been applied, whenever a special family of mss. clearly separated from the rest - for example by a large omission - has been investigated separately before turning to the rest of the mss. But this type of localizing was occasional, not systematic. The systematic application of localizing aims at the separate construction of branches or sections of branches which are as small as possible, but at the same time large enough to yield the connection of the complete stemma. No section should contain paths touching more than two preserved mss. : small sections and short paths will contain only few irregularities such as corrections, contaminations or identical but independent faults, a fact which makes local methods superior to global ones. If two neighbouring local sections overlap only in a single ms., the combination of the separate constructions will be consistent. Most convenient are sections rooted in a single preserved ms. and including the paths to its first preserved successors, as shown in the following figure :



But how can the mss. of such a small and connected section be found, if the stemma is not known ? It is immediately clear that a terminal section is associated with a family of exactly one non-terminal and at least one terminal ms.; in a locally regular tradition, this family does not contain a smaller one with the same property. So a minimal set of mss. - all characterized by common readings not found elsewhere and all but one terminal - is very likely to be the family of a terminal section. When this section has been constructed, the terminal mss. of the section are eliminated and a new terminal family is looked for, etc. For constructing a single section, a generalized version of the method of Froger has been used here(1); this version can handle certain types of contamination.

The first examples are related to an anonymous Greek text on musical theory preserved in 19 mss. and published as "Anonyma de musica scripta Bellermanniana"(2); traditional methods, based on a complete collation and leaving out of account only slight prosodical and orthographical deviations, have led to the following stemma :



With great certainty, this stemma can be regarded as correct; only a few readings might be due to a hyparchetype γ over 18 and 19, and the parallel tradition of Aristides Quintilianus, whose books on music precede our text in the mss., supports the introduction of an hyparchetype β over 16 and 18.

For the automatic stemma construction, the data of the collation have been put into the following form :

A	B	C	D E F
7	0000000000000001110	T4EXNH MOUSIK5HS	000.01.0
10	1111111111111110000	OM.	000.01.1
340	0000000000000000001	OM. SS. 1 - 28	000.01.2
7	1111111111111101110	A34IDE	001.01.0
8	0000000000000010000	A3I D4E	001.01.1
7	0000000000000001110	D4IXRONOS -	001.02.0
6	1111111111111110000	D4IXRONOS =	001.02.1
7	1110111111111001110	***	001.03.0
10	0001000000000110000	OM.	001.03.1

Each variant is associated with a weight in column A and with an incidence vector of the related mss. in column B. Columns D, E and F give the reference : paragraph, number of the variant place in the paragraph, and number of the variant at this place; variant number 0 indicates the readings regarded as original; the corresponding constellations, i.e. the sets of mss. defined by the incidence vectors, will not be used for the first automatic constructions.

The weights have been chosen according to the following rules :

1.-	orthographical faults, easily corrigible	2
2.-	variants :	
	- easily corrigible (non-existing words or forms)	4
	- possibly corrigible (faults in grammar or meaning)	6
	- hardly corrigible (adequate in grammar and meaning)	8
3.-	doublets :	
	- up to 4 words	4
	- more than 4 words	6
4.-	insertions :	
	- 1 or 2 words, not fitting to the meaning	6
	- 1 or 2 words, fitting to the meaning	8

ANNEXE

sur le calcul du modèle de Poisson

Il y a avantage, pour appliquer ce modèle, à choisir la longueur des tranches de façon que le nombre moyen d'occurrences par tranche se situe entre 0,1 et 5, mais de préférence assez près de 1.

Ainsi, pour 626 pages, nous avons 265 occurrences, soit 0,4233 en moyenne par page, ce qui est satisfaisant. Mettons en mémoire ce nombre, que nous représenterons par *m*. On forme ensuite le nombre *e* (qui est 2,7182818 . . .); si la calculatrice ne le donne pas directement, on peut l'obtenir en formant l'antilog naturel de 1 :

1 INV lnx

On calcule ensuite e^{-m}, en élevant *e* à la puissance *m* (rappel de mémoire), puis en faisant l'inverse de ce résultat :

y^x RCL = 1/x

Cette expression (dans notre exemple, 0,65486728) est la probabilité pour qu'une tranche ait 0 occurrence; multipliant cette probabilité par le nombre de pages, on obtient 626 x 0,6549 = 409,95, nombre théorique des pages à fréquence 0. Les nombres suivants du modèle s'obtiennent en multipliant le premier résultat par *m*, puis le second par *m*/2, le troisième par *m*/3, et ainsi de suite.

409,95 x 0,4233 = 173,54

173,54 x 0,4233/2 = 36,73, etc. . .

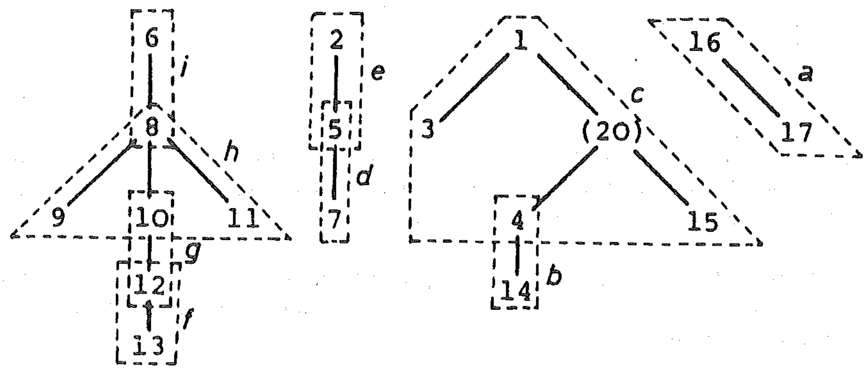
If the first two constellations, for example, could be accumulated, a fuzzy set would arise, with a lower degree of membership for ms. 3. The linking function of ms. 3 is present in the above set of constellations, but it seems that the special role of such transition forms cannot be fully exhausted by constellation based methods, not even with fuzzy sets, and this applies equally to paleographical peculiarities which can be understood as transition forms.

Also the age of the mss., though being vague in many cases, can exclude certain connections, but it is not used in our constellation based methods. So the description of the data shows that some information exploited by conventional methods is not fully used by local genealogical algorithms, at least in their present state.

In a last preparatory step, equal constellations and their weights are accumulated, so that each constellation type occurs only once, but with the sum of the weights of all of its occurrences. The accumulated weights are used in the following way : if several constellations are candidates for the set of mss. to be considered locally, the constellation with maximum weight will be chosen. Furthermore, constellations which do not reach a certain minimal weight are not taken account of at all, neither for localizing nor for constructing. The minimal weight may be varied systematically, in order to obtain a whole series of stemmata.

The problem of how to determine the terminal mss. will be excluded for the stemma constructions related to the Greek text considered : we shall regard those mss. as terminal which have been shown to be terminal by the conventional method. This allows us to locate the origin of construction faults in the proper algorithm.

Now to the results of automatic construction. Original weights yield a sequence *a, b, . . ., i* of local constructions as follows :

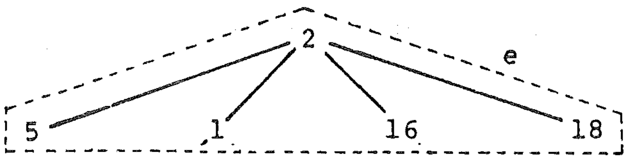


The single branches correspond well to those of the known stemma. Only few deviations must be analysed.

In construction *b*, ms. 15 is excluded, because - on account of corrections in 15 - both constellations [4,14,15] and [4,14] are candidates for localizing, and the smaller candidate is chosen. Only with high minimal weights - 27 and 30 - ms. 4 will occupy the place of ms. 20; obviously the corrections in ms. 15 do not reach these weights. With low minimal weights, ms. 15 cannot become a successor of ms. 4, since ms. 4 is treated as terminal after the elimination of ms. 14. 20 has been put in parentheses indicating that it has been introduced as a hypothetical lost ms. by the algorithm; the numbering of hypothetical mss. is arbitrary, so that the same number may represent different lost mss. in alternative stemmata.

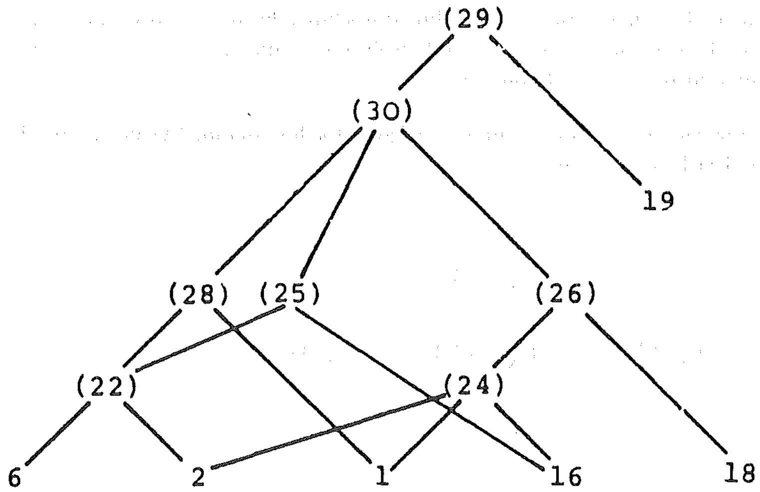
After constructions *a* and *b*, constellation [1,3,4,15] is - by its weight - the first candidate for the next local construction. This means that corrections in 2 and further alterations of faults of 1 surpass the weights of the concurrent candidates as constellations [5,7] or [12,13] ; it does not mean that the faults of 1 taken over in 2 have a lower weight than corrections and further alterations, but 2 is not yet terminal at this stage, so that [1,2,3,4,15] is no candidate for local construction and [1,3,4,15] is chosen. After the construction, mss. 3, 4 and 15 are eliminated and 1 is treated as terminal, so that 2 cannot be appended to 1 in a later step.

Analogously, ms. 6 cannot be appended to ms. 2.
The described sequence of local constructions remains the same for minimal weights 3,6,9, . . .,18. Minimal weights 21 and 24 modify construction *e* as follows :



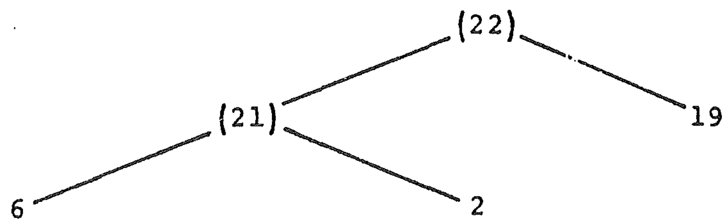
This is obviously wrong, since ms. 1 is older than ms. 2. The reason for this mistake is that the search for a minimal localizing candidate was not successful, constellation [2,5] does not reach the necessary weight. So constellation [1,2,5,16,18] is chosen.

At this point of the procedure - for minimal weights up to 18 - all mss. except 1,2,6,16,18,19 have been eliminated, and the remaining ones are terminal now. For minimal weight 18, the last local construction yields the following result :



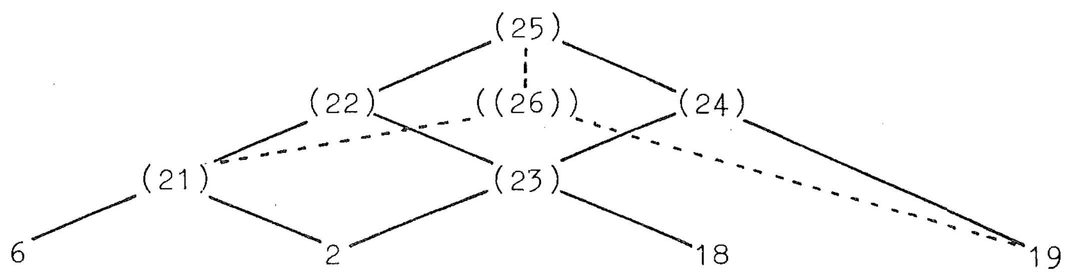
Since 2 and 6 did not obtain their preserved predecessors before (1 and 2), there are too many mss. remaining for the last local construction. This raises the chance that corrections and parallel faults result in irregular constellations, and all of these will introduce lost mss.; in fact this has happened, only ms. 26 can be identified with the hyparchetype β. For lower minimal weights, even more lost mss. are introduced.

For minimal weights 21 to 30, the last local construction yields the following result :



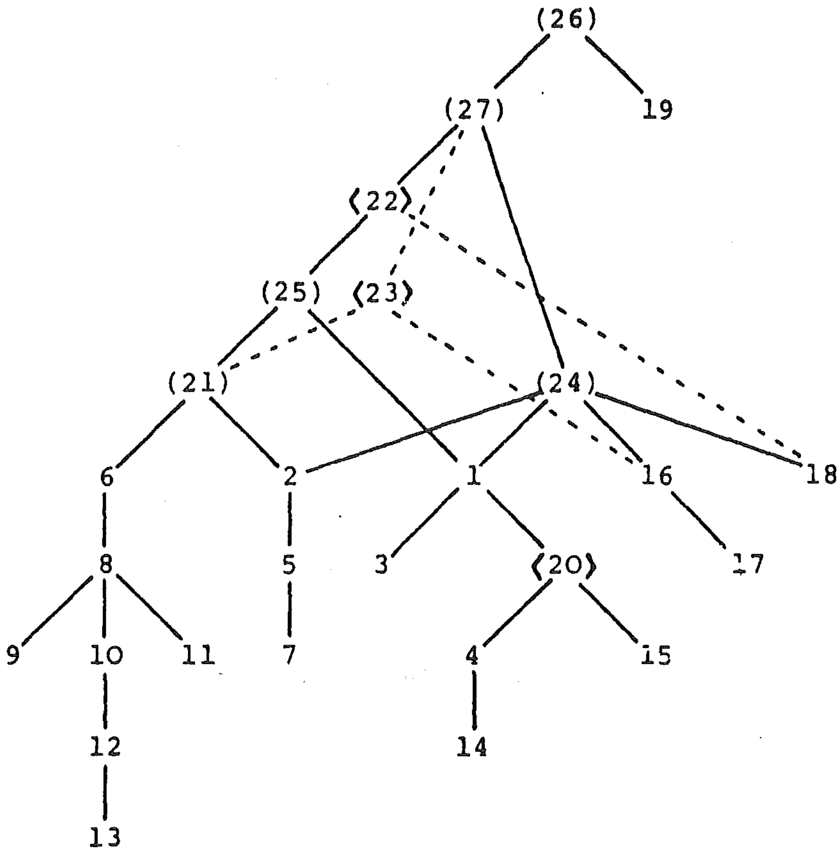
This is better with respect to the number of lost mss., but unfortunately ms. 2 had become the predecessor of ms. 1 before. So it seems preferable to use relatively low minimal weights and to restrict the introduction of lost mss. in the last local construction.

Cubic weights yield the same stemma for high minimal weights; for low minimal weights more lost mss. are introduced in the last local construction :



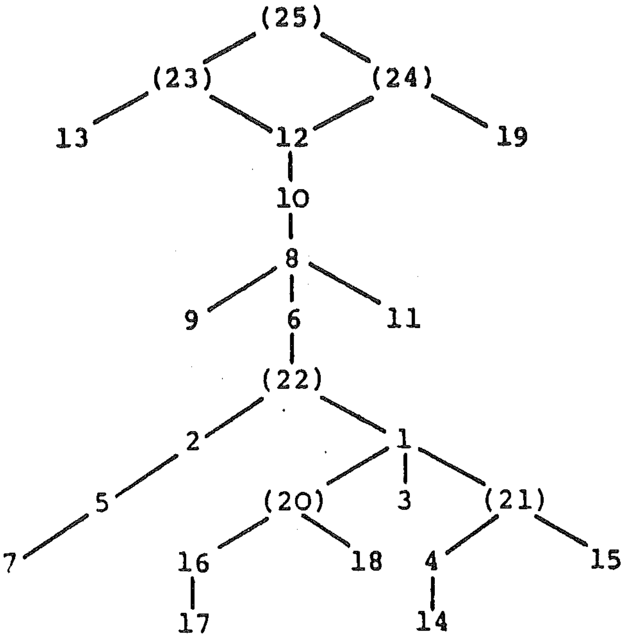
Ms. 26 is only present for minimal weight 10.

The next stemma is based on standard weights. Minimal weights 15 and 20 yield the following result :

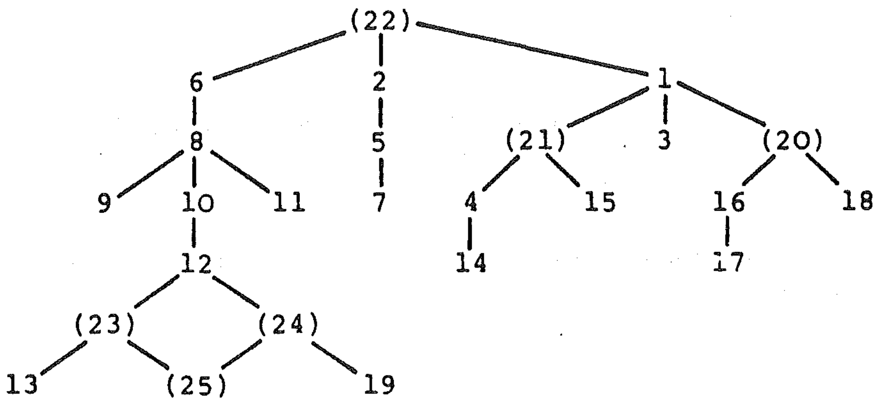


With minimal weight 20, mss. 22 and 23 and the broken lines disappear. It is surprising that the single branches depending on mss. 1,2,6,16 are exactly the same as with individual weights. It must be concluded that the text is long enough (417 variant places) to balance most differences of weights.

A last stemma construction related to the same text shows the influence of using constellations defined by original readings along with constellations defined by faults. With minimal weight 18, the following stemma results :



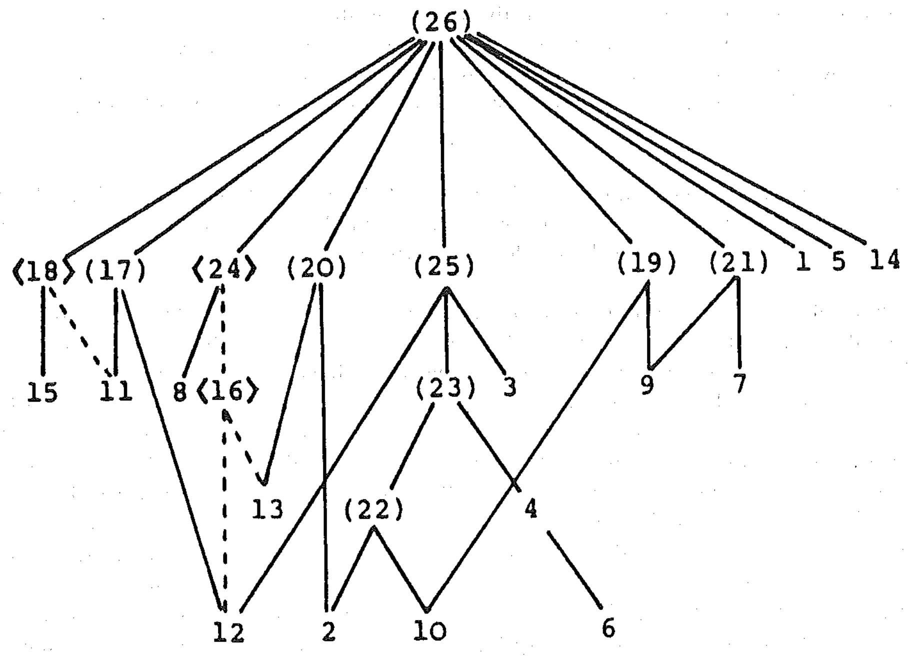
This is rather different from the correct family tree, but when the stemma is suspended at ms. 22, it looks more familiar :



In fact, if original readings and faults are used in the same way, the orientation of the stemma becomes arbitrary, so that the correct orientation must be found afterwards. But the left hand branch of the stemma shows that also certain other distortions may arise which cannot be resolved by changing the orientation. In any case it seems preferable to avoid constellations defined by original readings.

Before passing to another text, we can, on the whole, state that the results of the algorithm - despite of a few inherent disadvantages - come close to the stemma found by conventional methods, if the introduction of lost mss. is restricted in the last local construction.

A last example is related to a medieval Latin text, the Tobias of Mathew of Vendôme. Of the 106 mss. preserving this text, only those 15 are considered here, which F. Munari has chosen as the base for his forthcoming edition. The choice has proved to be good : high weights for all singleton constellations show that hardly any of the 15 mss. can be assumed to depend on another one in the same set, but that all mss., except perhaps ms. 4, must be treated as terminal. This means that we are almost at once in the situation of the last local construction. On the other hand, the situation seems to be even worse for conventional editing, since the number of constellations is very high in this case (1,111), and many of them would be mutually exclusive in a regular tradition. There are, for example, 69 constellations of exactly 2 mss., almost 2/3 of the largest possible number, which is $\binom{15}{2} = 105$. For the Greek text considered above, we had only 10 constellations of 2 mss., about 6 % of the largest possible number of $\binom{19}{2} = 171$. So it is rather evident that the tradition of the Tobias is heavily contaminated, and it is not surprising that human mind has difficulties in handling the great number of constellations and to find a structure in this host. This, however, is not a difficulty for computers, so that the algorithmic approach, despite of the bad starting conditions, is helpful in this case. Whereas conventional methods failed to find a stemma, automatic construction yields the following one :



As a conclusion of this comparison of conventional methods of textual criticism and local genealogical algorithms, we can state that at present, the latter seem to be almost as good as the first and in single instances even superior. A further improvement of automatic stemma construction seems possible, for example by a refinement of localizing methods or by a redefinition of terminal mss. before the last local construction. The influence of different weighting schemes is negligible, if the number of variant places is high enough. Constellations defined by original readings mainly affect the orientation of the stemma, but they may also cause single distortions.

REFERENCES

(1) D. Najock, A Genealogical Algorithm Operating on Minimal Intersecting Subgraphs, in : La Pratique des Ordinateurs dans la Critique des Textes (Colloques Internationaux du CNRS, n. 579), Paris, 1979, p. 211-225.

(2) Anonyma de musica scripta Bellermanniana, ed. D. Najock, Leipzig, 1975.