Hierarchical Clustering Methods for Asymmetric Networks

F.Mémoli.

Joint with G.Carlsson, A. Ribeiro, and S.Segarra.

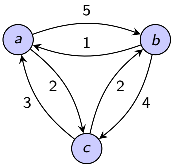
http://arxiv.org/abs/1301.7724,

Sept. 2014.

Some basic concepts

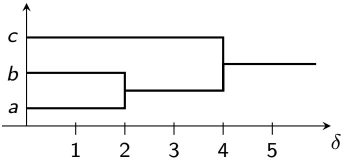
Asymmetric network

- Weighted and directed
- $N = (X, A_X)$
  - $\Rightarrow$ Node set $X = \{a, b, c\}$
  - $\Rightarrow$ Weights in A_X represent dissimilarities



Hierarchical clustering

- Traditional clustering $\Rightarrow$ Partition of node set $\Rightarrow$ e.g. $\{\{a, b\}, \{c\}\}$
- Hierarchical clustering $\Rightarrow$ Nested collection of partitions $\Rightarrow$ Dendrogram



$$D(1) = \{\{a\}, \{b\}, \{c\}\}$$

$$D(3) = \{\{a, b\}, \{c\}\}$$

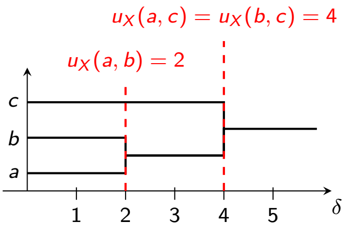
$$D(5) = \{\{a, b, c\}\}$$

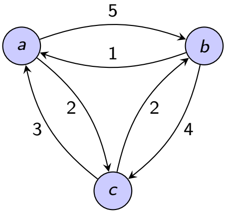
Method

- Map $\mathcal{H} : \mathcal{N} \to \mathcal{D}$ from the space of networks to the space of dendrograms

Dendrograms as ultrametrics

- Dendrograms can be interpreted as discrete ultrametrics on node set X
- Ultrametric associated with dendrogram $\Rightarrow u_X(x,x') = \min_{\delta} \left\{ \delta \mid x \sim x' \right\}$
- Resolution at which x and x' are joined together



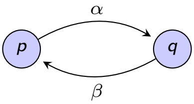


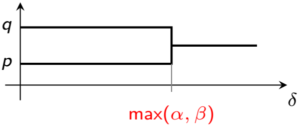
- Satisfy strong triangle ineq. $\Rightarrow u_X(x,x') \leq \max\Big(u_X(x,x''), u_X(x'',x')\Big)$
- We can reinterpret clustering methods $\mathcal{H} : (X, A_X) \to (X, U_X)$
- Which methods $\mathcal{H}$ are reasonable? Impose two axioms [Carlsson et al '13]

Axioms of value and transformation

Axiom of Value

- Fixes the behavior of clustering methods in two-node networks

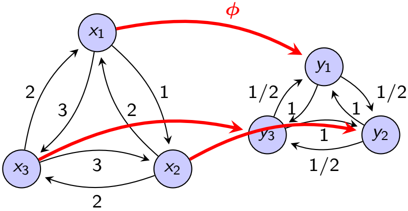


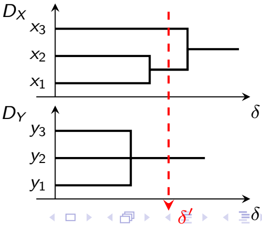


- To form a cluster nodes should be able to influence each other

Axiom of Transformation

- If nodes are 'closer', they have to cluster at earlier resolutions

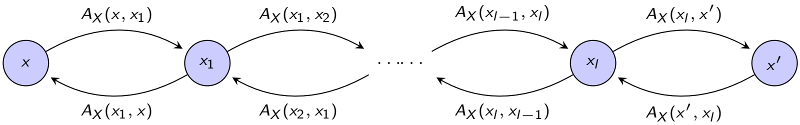




Reciprocal clustering

- $C(x, x')$ is a chain linking x and x' $\Rightarrow C(x, x') = [x = x_0, x_1, \ldots, x_l = x']$
- Chain cost is the maximum dissimilarity encountered in the chain

$$\Rightarrow \max_i \max\Big(A_X(x_i, x_{i+1}), A_X(x_{i+1}, x_i)\Big)$$



- x, x' clustered at resolution δ if they can be linked by paying less than δ

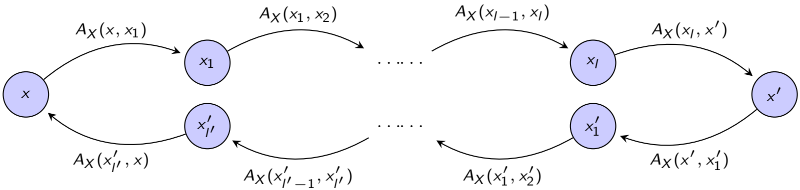
$$u_X^R(x, x') = \min_{C(x,x')} \Big[\max_i \max\Big(A_X(x_i, x_{i+1}), A_X(x_{i+1}, x_i)\Big)\Big]$$

- Single linkage on symmetrized network $(X, \max(A_X, A_X^T))$

Nonreciprocal clustering

- In reciprocal path, x and x' connected back and forth through same chain
- Nonreciprocal clustering allows different chains

$$u_X^{NR}(x,x') = \max\left[\min_{C(x,x')}\left(\max_i A_X(x_i,x_{i+1})\right),\ \min_{C(x',x)}\left(\max_i A_X(x_i,x_{i+1})\right)\right]$$



- Both methods satisfy axioms of value and transformation

Extremal methods

- Reciprocal and nonreciprocal bound every other admissible method

Theorem
Consider a clustering method $\mathcal{H}$ satisfying the axioms of value and transformation. For any network $N_X = (X, A_X)$ denote as u_X the outcome of $\mathcal{H}$ applied to N_X. Then

$$u_X^{NR}(x,x') \ \leq \ u_X(x,x') \ \leq \ u_X^{R}(x,x').$$

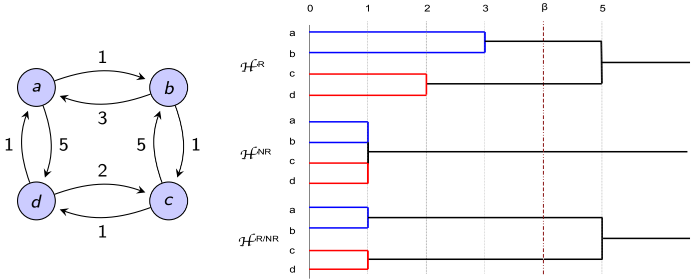
- No method yields ultrametrics smaller than $\mathcal{H}^{NR}$ or larger than $\mathcal{H}^{R}$
- In particular, if N_X is symmetric u_X^{NR} and u_X^{R} coincide
  $\Rightarrow$ Method is unique $\Rightarrow$ Single linkage
  $\Rightarrow$ Generalizes [Carlsson and Mémoli '10]
- **Existence of clustering methods between reciprocal and nonreciprocal**
- **Implementation of reciprocal, nonreciprocal and intermediate methods**

Grafting methods

- Constructed by cutting one dendrogram

  $\Rightarrow$ Pasting corresponding branches on another dendrogram

$$u_X^{\mathrm{R/NR}}(x,x';\beta) := \begin{cases} u_X^{\mathrm{NR}}(x,x'), & \text{if } u_X^{\mathrm{R}}(x,x') \leq \beta, \\ u_X^{\mathrm{R}}(x,x'), & \text{if } u_X^{\mathrm{R}}(x,x') > \beta. \end{cases}$$



- Cycles allowed for close nodes, not allowed for far away nodes

  $\Rightarrow$ Parameter β controls notions of close and far away

- From the four grafting possibilities, this is the only valid one

Convex combination methods

- Combine two admissible methods by combining output ultrametrics
  - $\Rightarrow$ Convex combinations of ultrametrics are NOT ultrametrics
  - $\Rightarrow$ They are symmetric $\Rightarrow$ One valid clustering method
- To combine admissible methods $\mathcal{H}^1$ and $\mathcal{H}^2$ into a new method $\mathcal{H}^{12}_\theta$
  - $\Rightarrow$ Given a network, combine the output ultrametrics of $\mathcal{H}^1$ and $\mathcal{H}^2$

$$A_X^{12}(x, x'; \theta) := \theta\, u_X^1(x, x') + (1-\theta)\, u_X^2(x, x')$$

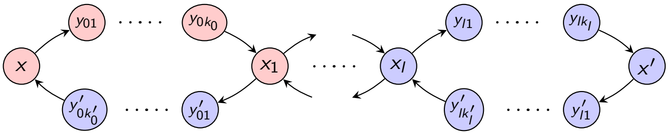
  - $\Rightarrow$ Cluster $(X, A_X^{12}(\theta))$ with any admissible clustering method
  - $\Rightarrow$ Since network is symmetric $\Rightarrow$ Only one admissible method
  - $\Rightarrow$ Single linkage

$$u_X^{12}(x, x'; \theta) := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} A_X^{12}(x_i, x_{i+1}; \theta)$$

- Parameter θ controls the relative weight of $\mathcal{H}^1$ in the method combination

Semi-reciprocal methods

- Reciprocal allows no cycles, nonreciprocal allows cycles of arbitrary length
- Allow cycles up to a maximum length $\Rightarrow$ semi-reciprocal clustering



- Secondary chains of maximum node length t $\Rightarrow \mathcal{H}^{\mathrm{SR}(t)}$
- $\mathcal{H}^{\mathrm{SR}(2)} \equiv \mathcal{H}^{\mathrm{R}}$ and $\mathcal{H}^{\mathrm{SR}(t)} \equiv \mathcal{H}^{\mathrm{NR}}$ for $t \geq n$
- Find optimal secondary chains between every pair of nodes

$$A_X^{\mathrm{SR}(t)}(x,x') := \min_{C_t(x,x')} \max_{k | x_k \in C_t(x,x')} A_X(x_k, x_{k+1})$$

- Concatenate secondary chains optimally

$$u_X^{\mathrm{SR}(t)}(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \max\left(A_X^{\mathrm{SR}(t)}(x_i, x_{i+1}), A_X^{\mathrm{SR}(t)}(x_{i+1}, x_i)\right)$$

- Parameter t controls cyclic propagation of influence

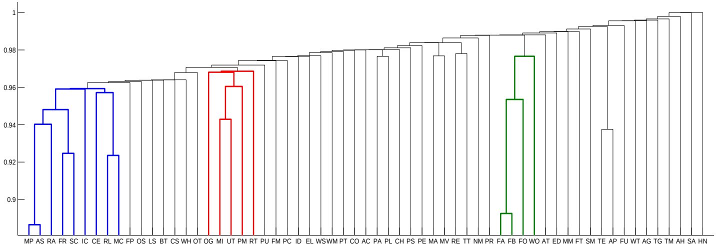
Economic sectors table

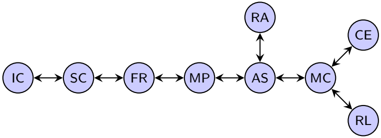
Bureau of Economic Analysis, Department of Commerce. Yearly publication. 61 Sectors.

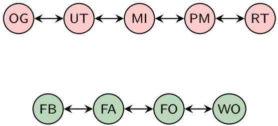
Code	Industrial Sector	Code	Industrial Sector
FA	Farms	TT	Truck transportation
FO	Forestry, fishing, and related activities	TG	Transit and ground passenger transportation
OG	Oil and gas extraction	PT	Pipeline transportation
MI	Mining, except oil and gas	OT	Other transportation and support activities
SM	Support activities for mining	WS	Warehousing and storage
UT	Utilities	PU	Publishing industries (includes software)
CO	Construction	PS	Motion picture and sound recording industries
WO	Wood products	BT	Broadcasting and telecommunications
NM	Nonmetallic mineral products	ID	Information and data processing services
PM	Primary metals	FR	Federal Reserve banks and credit intermediation
FM	Fabricated metal products	SC	Securities, commodity contracts, and investments
MA	Machinery	IC	Insurance carriers and related activities
CE	Computer and electronic products	FT	Funds, trusts, and other financial vehicles
EL	Electrical equipment, appliances, and components	RA	Real estate
MV	Motor vehicles, bodies and trailers, and parts	RL	Rental and leasing serv. and lessors of intang. assets
TM	Other transportation equipment	LS	Legal services
FU	Furniture and related products	CS	Computer systems design and related services
MM	Miscellaneous manufacturing	MP	Misc. professional, scientific, and technical services
FB	Food and beverage and tobacco products	MC	Management of companies and enterprises
TE	Textile mills and textile product mills	AS	Administrative and support services
AP	Apparel and leather and allied products	WM	Waste management and remediation services
PA	Paper products	ED	Educational services
PR	Printing and related support activities	AH	Ambulatory health care services
PC	Petroleum and coal products	HN	Hospitals and nursing and residential care facilities
CH	Chemical products	SA	Social assistance
PL	Plastics and rubber products	PE	Performing arts, spectator sports and museums
WH	Wholesale trade	AG	Amusements, gambling, and recreation industries
RE	Retail trade	AC	Accommodation
AT	Air transportation	FP	Food services and drinking places
RT	Rail transportation	OS	Other services, except government
WT	Water transportation		

Numerical examples: Reciprocal clustering

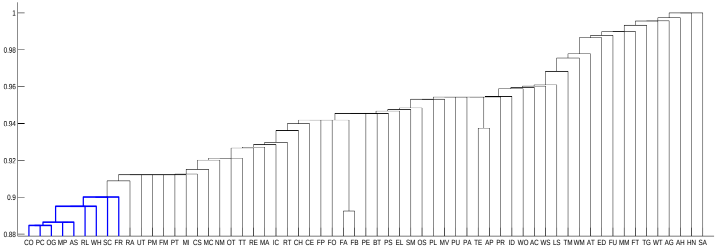
- Network of input-output interaction between economic sectors in U.S.

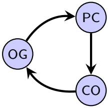


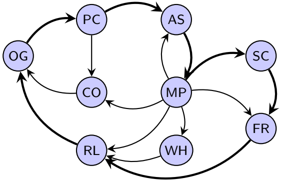




Numerical examples: Nonreciprocal clustering

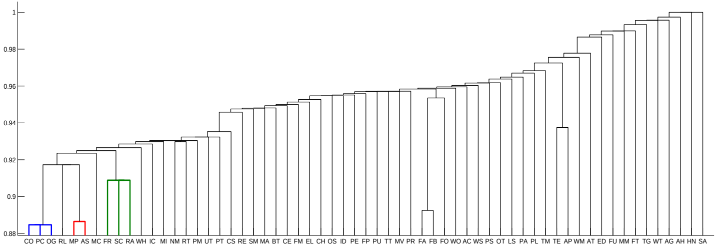


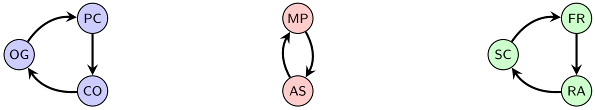




- Want to detect small cycles and avoid longer ones $\Rightarrow$ Semi-reciprocal

Numerical examples: Semi-reciprocal (t=3) clustering





Conclusion

- Expanded an axiomatic theory of hierarchical clustering in networks
- Developed admissible intermediate methods
  - $\Rightarrow$ Grafting $\Rightarrow$ Cut and paste branches of dendrograms
  - $\Rightarrow$ Convex combinations $\Rightarrow$ Combine output ultrametrics
  - $\Rightarrow$ Semi-reciprocal $\Rightarrow$ Limit cycle formation
- Presented natural framework for algorithmic development
  - $\Rightarrow$ Matrix operations in a min-max dioid algebra
  - $\Rightarrow$ Computation of chain costs

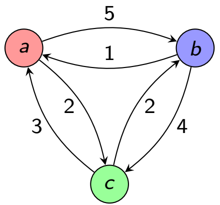
Algorithms: min-max dioid algebra

- Naturally understood in a dioid algebra $\Rightarrow$ min-max ($\oplus \to \min, \otimes \to \max$)

$$(2 \otimes 4) \oplus (5 \otimes 1) = 4 \oplus 5 = 4$$

- Dissimilarity function A_X can be reinterpreted as a $|X| \times |X|$ matrix
- $[A_X^k]_{i,j}$ is the minimum 'chain cost' of going from i to j in at most k hops
- $A_X = \begin{pmatrix} 0 & 5 & 2 \\ 1 & 0 & 4 \\ 3 & 2 & 0 \end{pmatrix}$

$$\begin{aligned} [A_X^2]_{b,c} &= 1 \otimes 2 \oplus 0 \otimes 4 \oplus 4 \otimes 0 \\ &= 2 \oplus 4 \oplus 4 \\ &= 2 \end{aligned}$$



- $A_X^2 = \begin{pmatrix} 0 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}$

Reciprocal and nonreciprocal algorithms

- If $|X| = n$, from the 'chain cost' interpretation $\Rightarrow A_X^{n-1} = A_X^n = \dots$

$$u_X^{\mathrm{R}} = \left(\max\left(A_X, A_X^T\right)\right)^{n-1}, \qquad u_X^{\mathrm{NR}} = \max\left(A_X^{n-1}, \left(A_X^T\right)^{n-1}\right)$$

- Complexity $O(n^4)$ in naive application since we need $n-1$ multiplications
  - $\Rightarrow$ Less if $A \to A^2 \to A^4 \to \cdots$
  - $\Rightarrow$ Sub-cubic dioid multiplication algorithms
- When A_X is symmetric, $u_X^{\mathrm{R}} = u_X^{\mathrm{NR}} = A_X^{n-1}$
- Methods are extremal in an algorithmic sense
  - $\Rightarrow$ Reciprocal: symmetrize then stabilize
  - $\Rightarrow$ Nonreciprocal: stabilize then symmetrize

Algorithms for intermediate clustering methods

Semi-reciprocal

- Compute finite power, symmetrize and stabilize

$$u_X^{\mathrm{SR}(t)} = \left(\max\left(A_X^{t-1}, (A_X^T)^{t-1}\right)\right)^{n-1}$$

- $t-1$ controls the allowable cycle length, $u_X^{\mathrm{SR}(2)} = u_X^{\mathrm{R}}$ and $u_X^{\mathrm{SR}(n)} = u_X^{\mathrm{NR}}$
- Natural intermediates from an algorithmic perspective

Grafting

- Combining matrices u_X^{R} and u_X^{NR}

$$u_X^{R/NR}(\beta) = u_X^{\mathrm{NR}} \circ \mathbb{I}\left\{u_X^{\mathrm{R}} \leq \beta\right\} + u_X^{\mathrm{R}} \circ \mathbb{I}\left\{u_X^{\mathrm{R}} > \beta\right\},$$

Convex combination

- Ultrametrics u_X^1 and u_X^2 are the outputs of two given methods

$$u_X^{12}(\theta) = \left(\theta\, u_X^1 + (1-\theta)\, u_X^2\right)^{n-1}$$

- Power $n-1$ clusters the symmetric matrix $\theta\, u_X^1 + (1-\theta)\, u_X^2$