Markov Clustering (MCL)

Michael Steyer

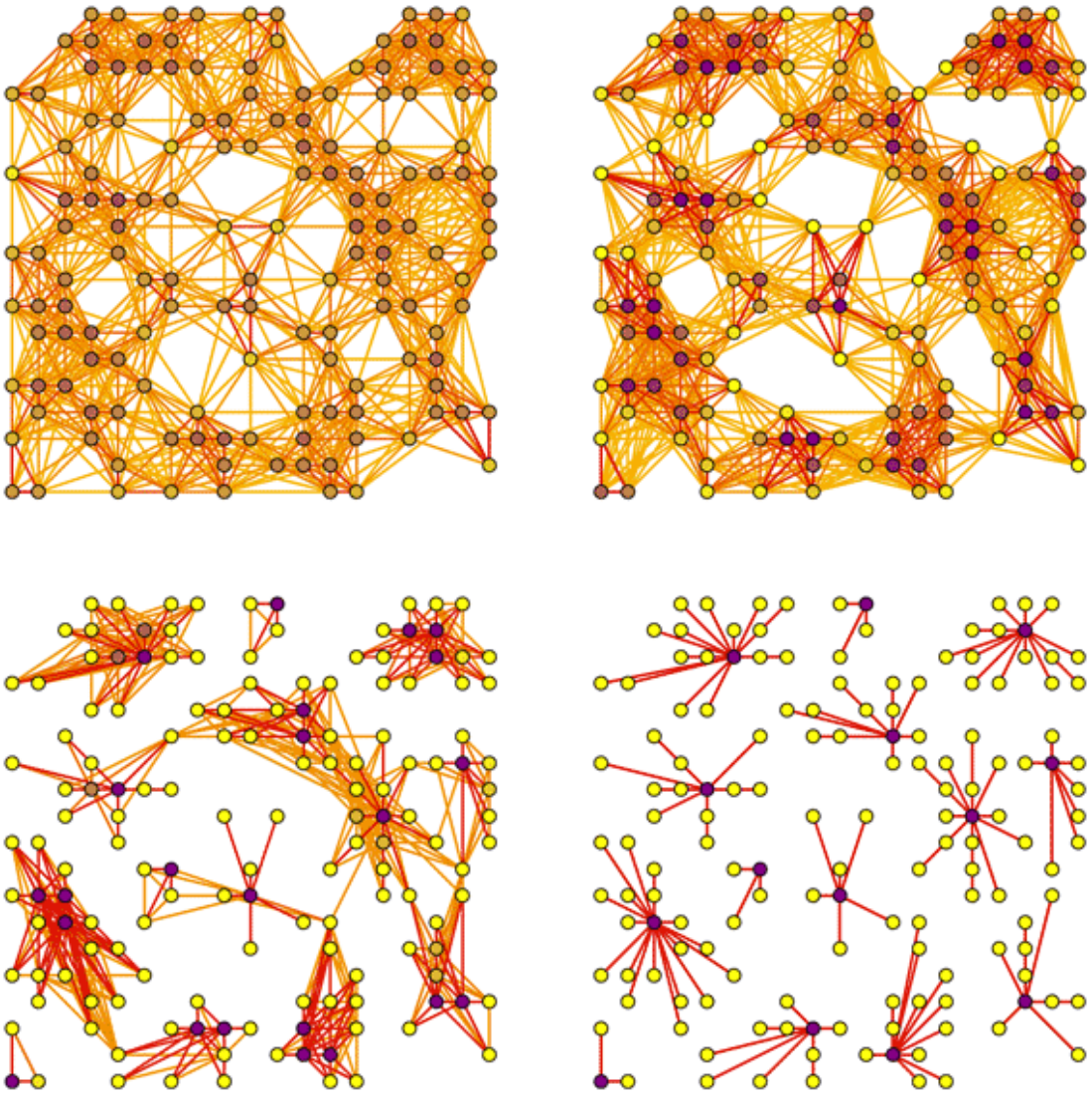
Universität Leipzig

20. Mai 2009

Agenda

1. Einführung
2. Algorithmus
3. Beispiel
4. Zusammenfassung

Clustering



Markov Clustering

- Entwickelt von Stijn van Dongen (Niederlande)

- Fluss Simulation clustering Methode
- Graph Clustering mittels „zufälligen Läufen“
- Lauf bleibt mit relativ großer Wkt. innerhalb des Clusters
- Nutzung der grundlegenden Eigenschaften von Clustern
  1. viele innere Kanten
  2. wenige äußere Kanten

Markov Clustering

- Erlaubt das Clustering von Graphen mit Kantengewichten
- Keinerlei Vorkenntnisse zur Cluster Struktur benötigt
- Weitere Vorteile
  1. sehr schnell
  2. sehr gut skalierbar
  3. lässt sich von Kanten zwischen verschiedenen Clustern nicht fehlleiten
  4. hat einen natürlichen Parameter für die Granularität
  5. klarer Mathematischer Zusammenhang zwischen den simulierten Prozessen und der Graphen-Struktur
  6. leicht verständlicher Algorithmus

Anwendungen

- Identification of ortholog groups
- protein complexes
- peer-to-peer node clustering
- image retrieval
- Word Sense Discrimination
- molecular pathway discovery
- structural domains
- ...

MCL - Funktionsweise

- Kein realer Lauf durch den Cluster!
- Aufbau einer Punkt zu Punkt Verbindung zwischen allen Knoten

MCL - Funktionsweise

Transformation des Graphen in eine Wahrscheinlichkeitsmatrix

Schrittweise weitere Transformation des Graphen (der Wkt.- Matrix)

Transformationsschritte:

1. expansion
2. inflation

Expansions- Schritt

Matrixmultiplikation
Spalten-Stochastische-Matrix (Wahrscheinlichkeitsmatrix) M (N x N)

$M_2 = M_1 * M_1$

Expansion steuert den Fluss innerhalb der Cluster

- rekombination
- multi-hop läufe

Effekt $\longrightarrow$ Clustering durch viele innere- und wenig äussere Kanten

Inflations- Schritt

Inflations- Schritt

Potenz- Koeffizient (Reelle Zahl) $r > 1$

Inflations- Schritt

Potenz- Koeffizient (Reelle Zahl) r > 1

Matrix M spaltenweise beeinflußt durch den
Potenz- Koeffizienten r $\widehat{=} \Gamma_r(M)$

Γ_r wird der Inflations- Operator mit Potenz- Koeffizienten r genannt

Inflations- Schritt

Potenz- Koeffizient (Reelle Zahl) r > 1

Matrix M spaltenweise beeinflußt durch den
Potenz- Koeffizienten r $\widehat{=}$ $\Gamma_r(M)$

Γ_r wird der Inflations- Operator mit Potenz- Koeffizienten r genannt

$\Sigma_{r,j}(M) \widehat{=}$ Addition der Werte in Spalte j aus M, nachdem die Werte mittels des Koeffizienten r potenziert wurden

$\Sigma_{r,j}(M) = \sum_{k=1}^{N} M_{kj}{}^{r}$

Inflations- Schritt

Potenz- Koeffizient (Reelle Zahl) r > 1

Matrix M spaltenweise beeinflußt durch den
Potenz- Koeffizienten r $\widehat{=}$ $\Gamma_r(M)$

Γ_r wird der Inflations- Operator mit Potenz- Koeffizienten r genannt

$\Sigma_{r,j}(M) \widehat{=}$ Addition der Werte in Spalte j aus M, nachdem die Werte mittels des Koeffizienten r potenziert wurden

$\Sigma_{r,j}(M) = \sum_{k=1}^{N} {M_{kj}}^r$

$\Gamma_r(M)$ wird elementweise definiert durch $\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)}$

Inflations- Schritt

Potenz- Koeffizient (Reelle Zahl) r > 1

Matrix M spaltenweise beeinflußt durch den
Potenz- Koeffizienten r $\widehat{=}$ $\Gamma_r(M)$

Γ_r wird der Inflations- Operator mit Potenz- Koeffizienten r genannt

$\Sigma_{r,j}(M) \widehat{=}$ Addition der Werte in Spalte j aus M, nachdem die Werte mittels des Koeffizienten r potenziert wurden

$\Sigma_{r,j}(M) = \sum_{k=1}^{N} {M_{kj}}^r$

$\Gamma_r(M)$ wird elementweise definiert durch $\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$

Inflations- Schritt

Inflation trennt die äußeren Kanten ab

- Verstärker
- Normierung

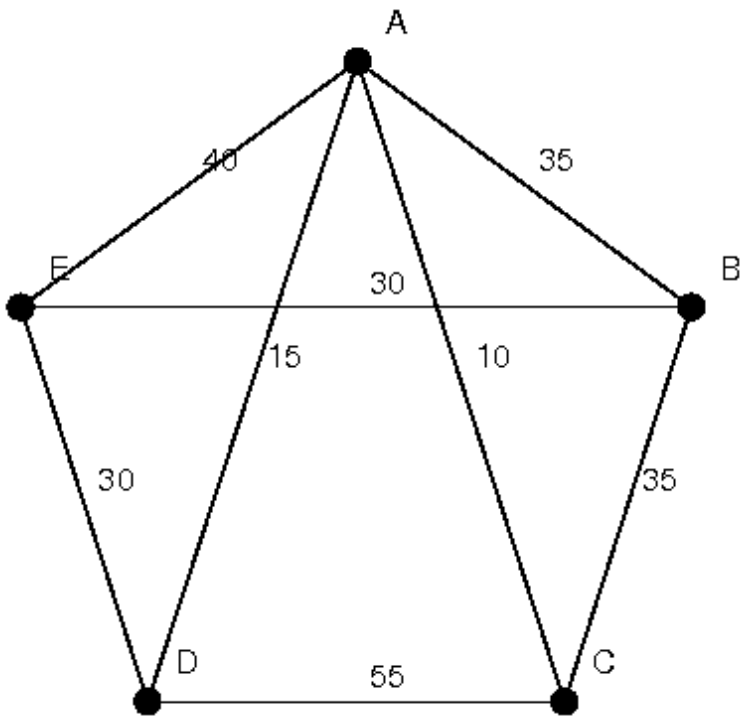
Ablauf des Algorithmus

$\mathbf{M} \leftarrow \mathbf{G}$
$\mathbf{M} \leftarrow \text{ADD_LOOPS}(\mathbf{M})$
$\mathbf{M} \leftarrow \text{NORMALIZE}(\mathbf{M})$

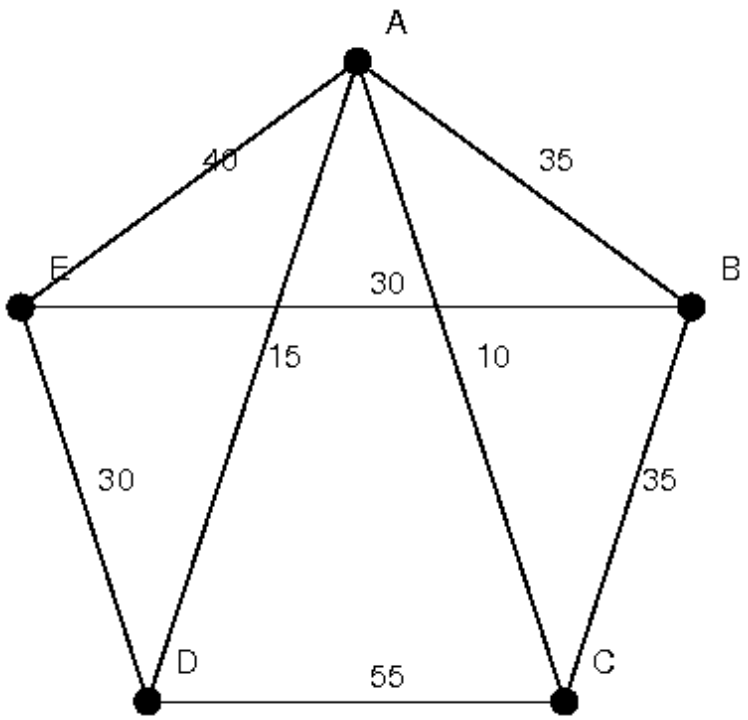
while diff **do**
  $M_2 \leftarrow M_1 * M_1$ #expansion
  $M_1 \leftarrow \Gamma_r(M_2)$ #inflation
  diff $\leftarrow$ DIFFERENCE(M_1, M_2)
end while

$\mathbf{M} \leftarrow \text{REMOVE_LOOPS}(\mathbf{M})$
$\mathbf{M} \leftarrow \text{NORMALIZE}(\mathbf{M})$
$\mathbf{G} \leftarrow \mathbf{M}$

Einfacher Graph



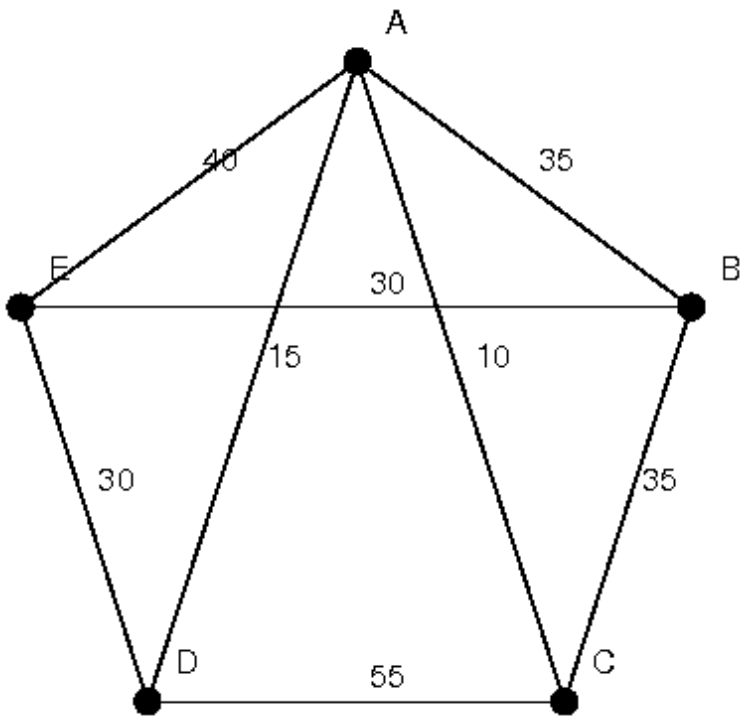
Einfacher Graph



Spalten-Stochastische-Matrix

$$\begin{pmatrix} 0.00 & 0.35 & 0.10 & 0.15 & 0.40 \\ 0.35 & 0.00 & 0.35 & 0.00 & 0.30 \\ 0.10 & 0.35 & 0.00 & 0.55 & 0.00 \\ 0.15 & 0.00 & 0.55 & 0.00 & 0.30 \\ 0.40 & 0.30 & 0.00 & 0.30 & 0.00 \end{pmatrix}$$

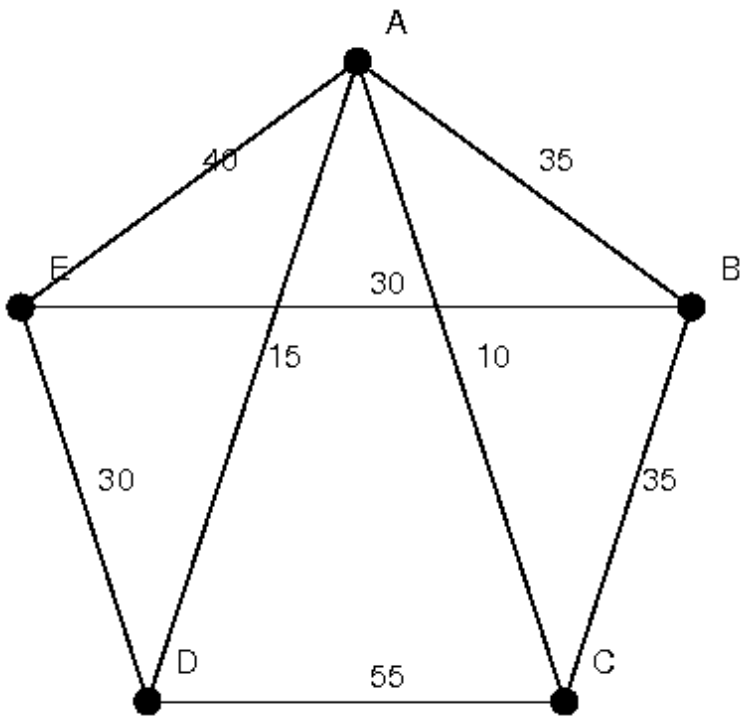
Einfacher Graph



~~Spalten-Stochastische~~-Matrix
incl. loops

$$\begin{pmatrix} 1.00 & 0.35 & 0.10 & 0.15 & 0.40 \\ 0.35 & 1.00 & 0.35 & 0.00 & 0.30 \\ 0.10 & 0.35 & 1.00 & 0.55 & 0.00 \\ 0.15 & 0.00 & 0.55 & 1.00 & 0.30 \\ 0.40 & 0.30 & 0.00 & 0.30 & 1.00 \end{pmatrix}$$

Einfacher Graph



Spalten-Stochastische-Matrix

incl. loops & normiert

$$\begin{pmatrix} 0.500 & 0.175 & 0.050 & 0.075 & 0.200 \\ 0.175 & 0.500 & 0.175 & 0.000 & 0.150 \\ 0.050 & 0.175 & 0.500 & 0.275 & 0.000 \\ 0.075 & 0.000 & 0.275 & 0.500 & 0.150 \\ 0.200 & 0.150 & 0.000 & 0.150 & 0.500 \end{pmatrix}$$

Expansions- Schritt

$$M_2 = M_1 * M_1$$

Expansions- Schritt

$$M_2 = M_1 * M_1$$

$$M_1 = \begin{pmatrix} 0.500 & 0.175 & 0.050 & 0.075 & 0.200 \\ 0.175 & 0.500 & 0.175 & 0.000 & 0.150 \\ 0.050 & 0.175 & 0.500 & 0.275 & 0.000 \\ 0.075 & 0.000 & 0.275 & 0.500 & 0.150 \\ 0.200 & 0.150 & 0.000 & 0.150 & 0.500 \end{pmatrix}$$

Expansions- Schritt

$$M_2 = M_1 * M_1$$

$$M_1 = \begin{pmatrix} 0.500 & 0.175 & 0.050 & 0.075 & 0.200 \\ 0.175 & 0.500 & 0.175 & 0.000 & 0.150 \\ 0.050 & 0.175 & 0.500 & 0.275 & 0.000 \\ 0.075 & 0.000 & 0.275 & 0.500 & 0.150 \\ 0.200 & 0.150 & 0.000 & 0.150 & 0.500 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 0.32875 & 0.21375 & 0.10125 & 0.11875 & 0.2375 \\ 0.21375 & 0.33375 & 0.18375 & 0.08375 & 0.1850 \\ 0.10125 & 0.18375 & 0.35875 & 0.27875 & 0.0775 \\ 0.11875 & 0.08375 & 0.27875 & 0.35375 & 0.1650 \\ 0.23750 & 0.18500 & 0.07750 & 0.16500 & 0.3350 \end{pmatrix}$$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Wähle $r > 1$ ▶ $r \rightarrow 2$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Wähle $r > 1$ ▶ $r \rightarrow 2$

$$M = \begin{pmatrix} 0.32875 & 0.21375 & 0.10125 & 0.11875 & 0.2375 \\ 0.21375 & 0.33375 & 0.18375 & 0.08375 & 0.1850 \\ 0.10125 & 0.18375 & 0.35875 & 0.27875 & 0.0775 \\ 0.11875 & 0.08375 & 0.27875 & 0.35375 & 0.1650 \\ 0.23750 & 0.18500 & 0.07750 & 0.16500 & 0.3350 \end{pmatrix}$$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Wähle $r > 1$ ▶ $r \rightarrow 2$

$$M = \begin{pmatrix} 0.32875 & 0.21375 & 0.10125 & 0.11875 & 0.2375 \\ 0.21375 & 0.33375 & 0.18375 & 0.08375 & 0.1850 \\ 0.10125 & 0.18375 & 0.35875 & 0.27875 & 0.0775 \\ 0.11875 & 0.08375 & 0.27875 & 0.35375 & 0.1650 \\ 0.23750 & 0.18500 & 0.07750 & 0.16500 & 0.3350 \end{pmatrix}$$

$$\Gamma_2(M_{11}) = \frac{M_{11}^2}{\sum_{k=1}^{5} {M_{k1}}^r}$$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Wähle $r > 1$ ▶ $r \rightarrow 2$

$$M = \begin{pmatrix} 0.32875 & 0.21375 & 0.10125 & 0.11875 & 0.2375 \\ 0.21375 & 0.33375 & 0.18375 & 0.08375 & 0.1850 \\ 0.10125 & 0.18375 & 0.35875 & 0.27875 & 0.0775 \\ 0.11875 & 0.08375 & 0.27875 & 0.35375 & 0.1650 \\ 0.23750 & 0.18500 & 0.07750 & 0.16500 & 0.3350 \end{pmatrix}$$

$$\Gamma_2(M_{11}) = \frac{M_{11}^2}{\sum_{k=1}^{5} {M_{k1}}^r} = \frac{0.32875^2}{0.32875^2+0.21375^2+0.10125^2+0.11875^2+0.23750^2}$$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Wähle $r > 1$ ▶ $r \rightarrow 2$

$$M = \begin{pmatrix} 0.32875 & 0.21375 & 0.10125 & 0.11875 & 0.2375 \\ 0.21375 & 0.33375 & 0.18375 & 0.08375 & 0.1850 \\ 0.10125 & 0.18375 & 0.35875 & 0.27875 & 0.0775 \\ 0.11875 & 0.08375 & 0.27875 & 0.35375 & 0.1650 \\ 0.23750 & 0.18500 & 0.07750 & 0.16500 & 0.3350 \end{pmatrix}$$

$$\Gamma_2(M_{11}) = \frac{M_{11}^2}{\sum_{k=1}^{5} {M_{k1}}^r} = \frac{0.32875^2}{0.32875^2+0.21375^2+0.10125^2+0.11875^2+0.23750^2}$$

$$= 0.460831734$$

Inflations- Schritt

$$\Gamma_r(M_{ij}) = \frac{M_{ij}^r}{\Sigma_{r,j}(M)} = \frac{M_{ij}^r}{\sum_{k=1}^{N} {M_{kj}}^r}$$

Wähle $r > 1 \blacktriangleright r \rightarrow 2$

$$\overline{M} = \begin{pmatrix} 0.46083 & 0.19687 & 0.03998 & 0.05614 & 0.23892 \\ 0.19482 & 0.47996 & 0.13167 & 0.02792 & 0.14497 \\ 0.04371 & 0.14548 & 0.50191 & 0.30934 & 0.02544 \\ 0.06013 & 0.03022 & 0.30302 & 0.49820 & 0.11532 \\ 0.24051 & 0.14747 & 0.02342 & 0.10839 & 0.47535 \end{pmatrix}$$

$$\Gamma_2(M_{11}) = \frac{M_{11}^2}{\sum_{k=1}^{5} {M_{k1}}^r} = \frac{0.32875^2}{0.32875^2+0.21375^2+0.10125^2+0.11875^2+0.23750^2}$$

$$= 0.460831734$$

Graph nach einer Iteration

Spalten-Stochastische-Matrix

$$\begin{pmatrix} 0.46083 & 0.19687 & 0.03998 & 0.05614 & 0.23892 \\ 0.19482 & 0.47996 & 0.13167 & 0.02792 & 0.14497 \\ 0.04371 & 0.14548 & 0.50191 & 0.30934 & 0.02544 \\ 0.06013 & 0.03022 & 0.30302 & 0.49820 & 0.11532 \\ 0.24051 & 0.14747 & 0.02342 & 0.10839 & 0.47535 \end{pmatrix}$$

Graph nach einer Iteration

~~Spalten-Stochastische~~-Matrix
ohne loops

$$\begin{pmatrix} 0.00000 & 0.19687 & 0.03998 & 0.05614 & 0.23892 \\ 0.19482 & 0.00000 & 0.13167 & 0.02792 & 0.14497 \\ 0.04371 & 0.14548 & 0.00000 & 0.30934 & 0.02544 \\ 0.06013 & 0.03022 & 0.30302 & 0.00000 & 0.11532 \\ 0.24051 & 0.14747 & 0.02342 & 0.10839 & 0.00000 \end{pmatrix}$$

Graph nach einer Iteration

Spalten-Stochastische-Matrix

ohne loops & normiert

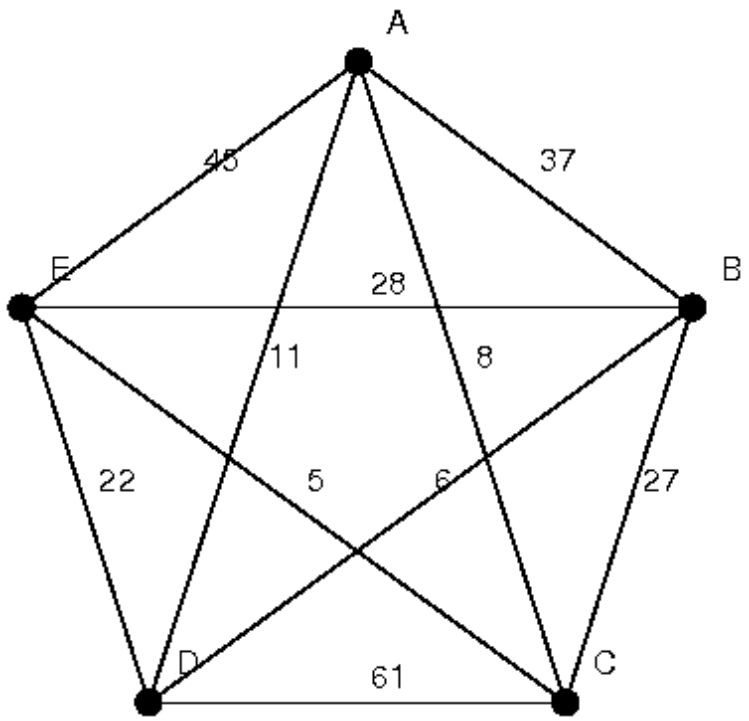
$$\begin{pmatrix} 0.00000 & 0.37857 & 0.08027 & 0.11188 & 0.45539 \\ 0.36133 & 0.00000 & 0.26435 & 0.05564 & 0.27632 \\ 0.08107 & 0.27975 & 0.00000 & 0.61647 & 0.04849 \\ 0.11152 & 0.05811 & 0.60836 & 0.00000 & 0.21980 \\ 0.44607 & 0.28357 & 0.04702 & 0.21601 & 0.00000 \end{pmatrix}$$

Graph nach einer Iteration

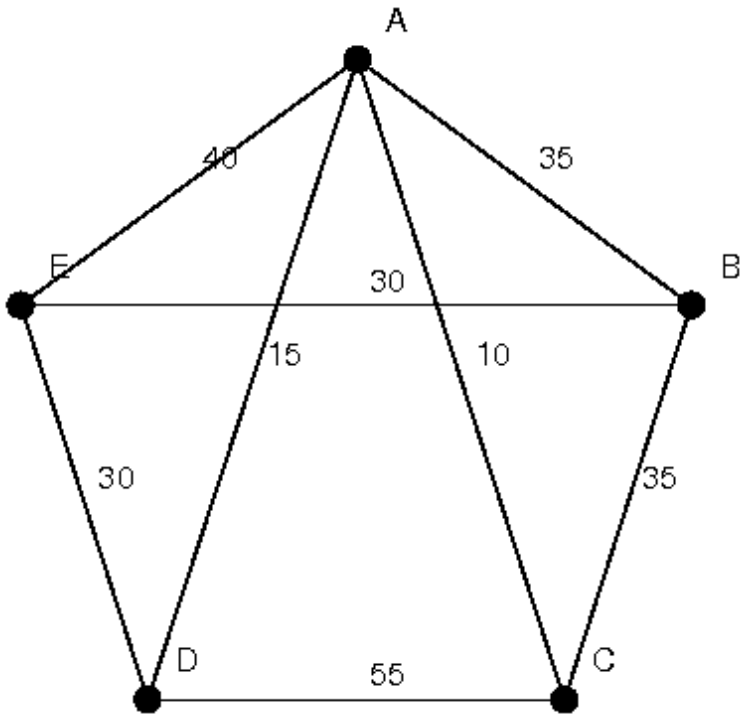
Spalten-Stochastische-Matrix

ohne loops & normiert

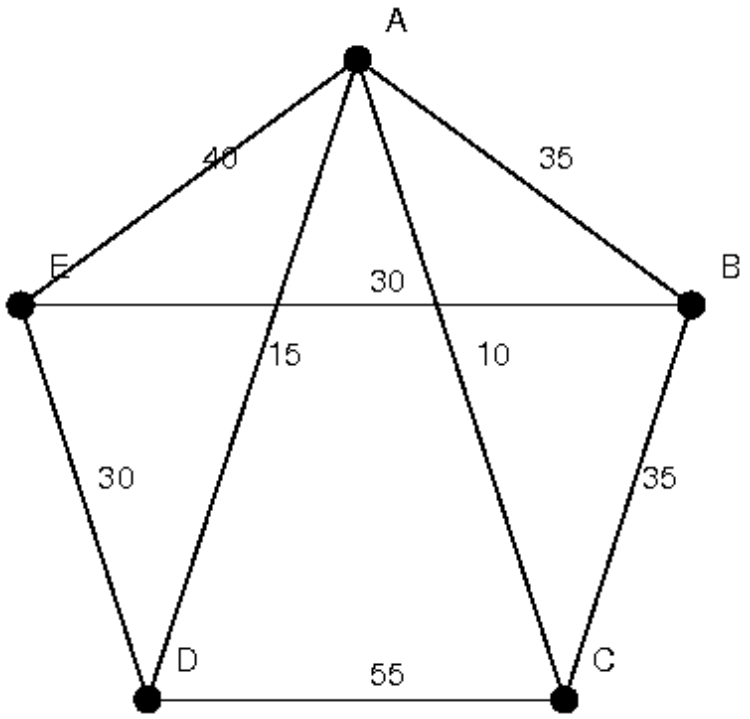
$$\begin{pmatrix} 0.00000 & 0.37857 & 0.08027 & 0.11188 & 0.45539 \\ 0.36133 & 0.00000 & 0.26435 & 0.05564 & 0.27632 \\ 0.08107 & 0.27975 & 0.00000 & 0.61647 & 0.04849 \\ 0.11152 & 0.05811 & 0.60836 & 0.00000 & 0.21980 \\ 0.44607 & 0.28357 & 0.04702 & 0.21601 & 0.00000 \end{pmatrix}$$

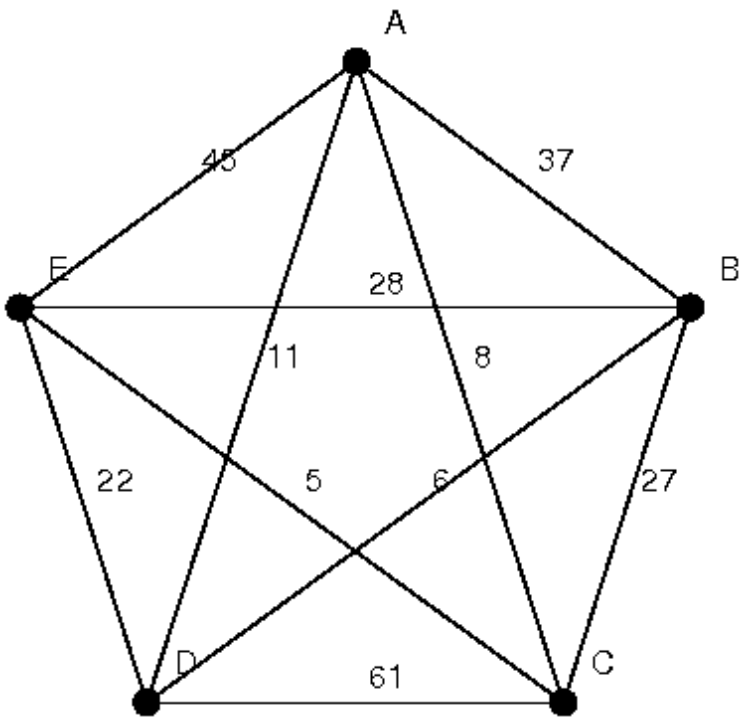


Vergleich nach einer Iteration



Vergleich nach einer Iteration





Komplexität

In der Praxis meist zwischen 3 und 10 Iterationen

Globale Konvergenz kann nicht garantiert werden,
mutmaßliche Konvergenz bei symmetrischem start- Graph

Zeit	$O(N^3)$
Speicher	$O(N^2)$

Parallelisierbarkeit

Expansion (Matrixmultiplikation)

- $O(N^3)$
- $O(N)$ (mit N^2 Prozessoren)

Inflation

- $3 * \sqrt{N}$ (mit $\sqrt{N}$ Prozessoren)

Parameter

Interessante Parameter

- r
- loops

Parameter

Interessante Parameter

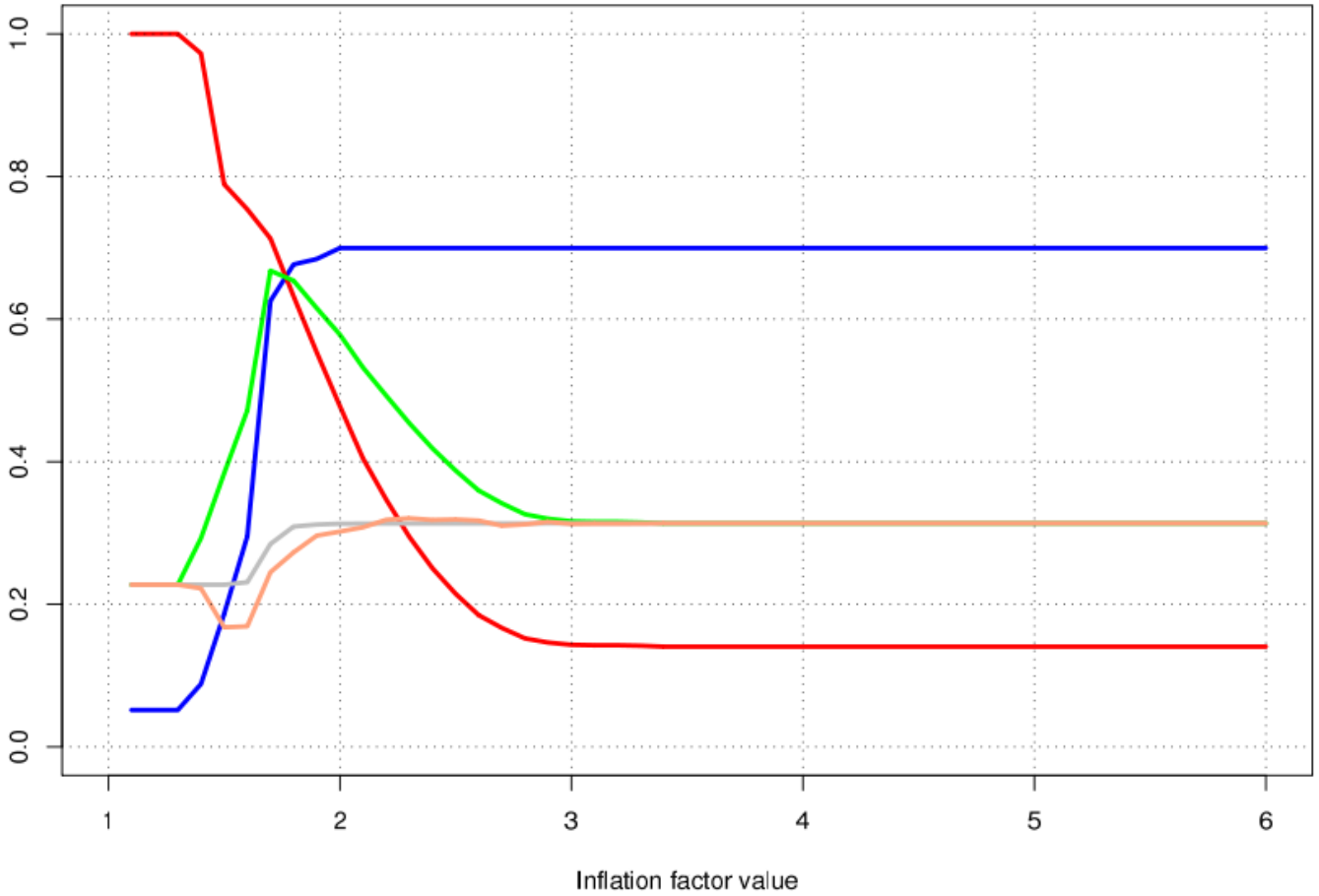
- $\uparrow r$ $\longrightarrow$ $\uparrow$ Granularität
- loops

Parameter

Interessante Parameter

- $\uparrow r \quad \longrightarrow \quad \uparrow$ Granularität
- $\uparrow loops \quad \longrightarrow \quad \uparrow$ Granularität

Parameter



Inflations-Parameter „r“

Blau → Feingranularität
Rot → Zusammenhang
Grün → Genauigkeit (∅)

Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

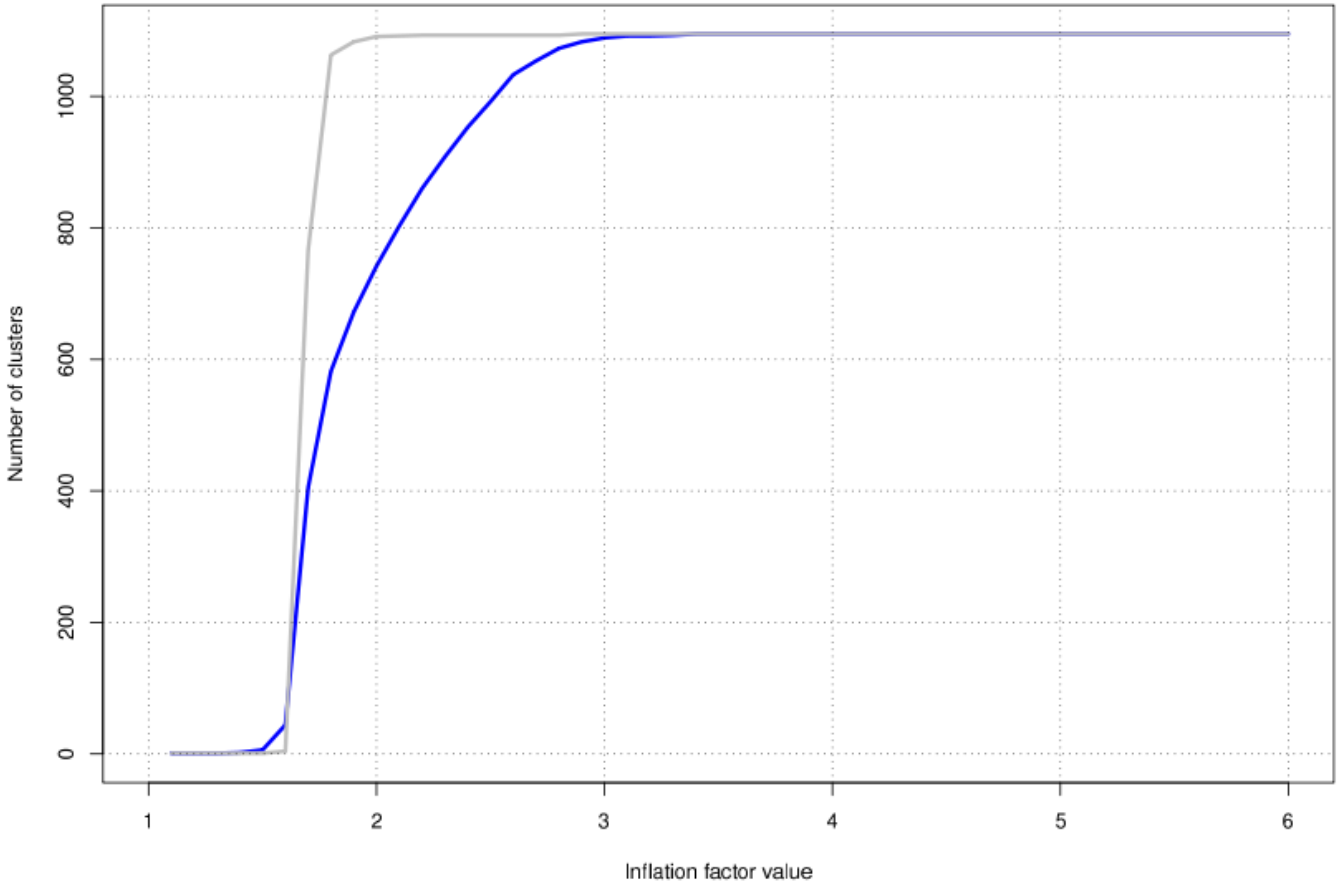
Parameter

Optimal values for MCL inflation parameter for the test and altered graphs

% removal\% addition	0	5	10	20	40	80	100
0	3.4	3.1	2.7	2.4	2	1.8	1.8
5	5.7	4	2.6	2	1.9	1.8	1.8
10	2.35	2.2	2.2	2.3	1.8	1.8	1.8
20	1.7	2.2	2.1	2	1.8	1.7	1.8
40	1.8	1.8	1.8	1.9	1.7	1.7	1.7
80	1.3	1.4	1.5	1.5	1.5	1.6	1.6

Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

Parameter

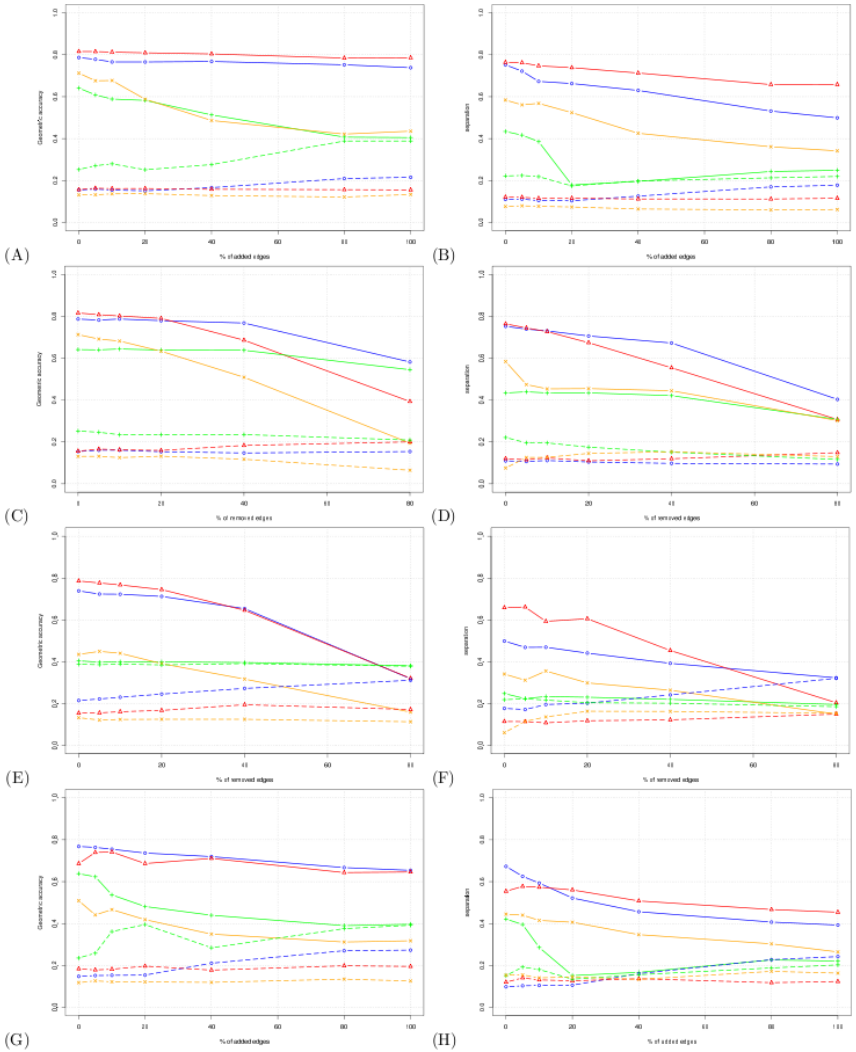


Inflations-Parameter „r“

Blau → Test- Graph (Paper)
Grau → Random- Graph

Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

Vergleich der Robustheit

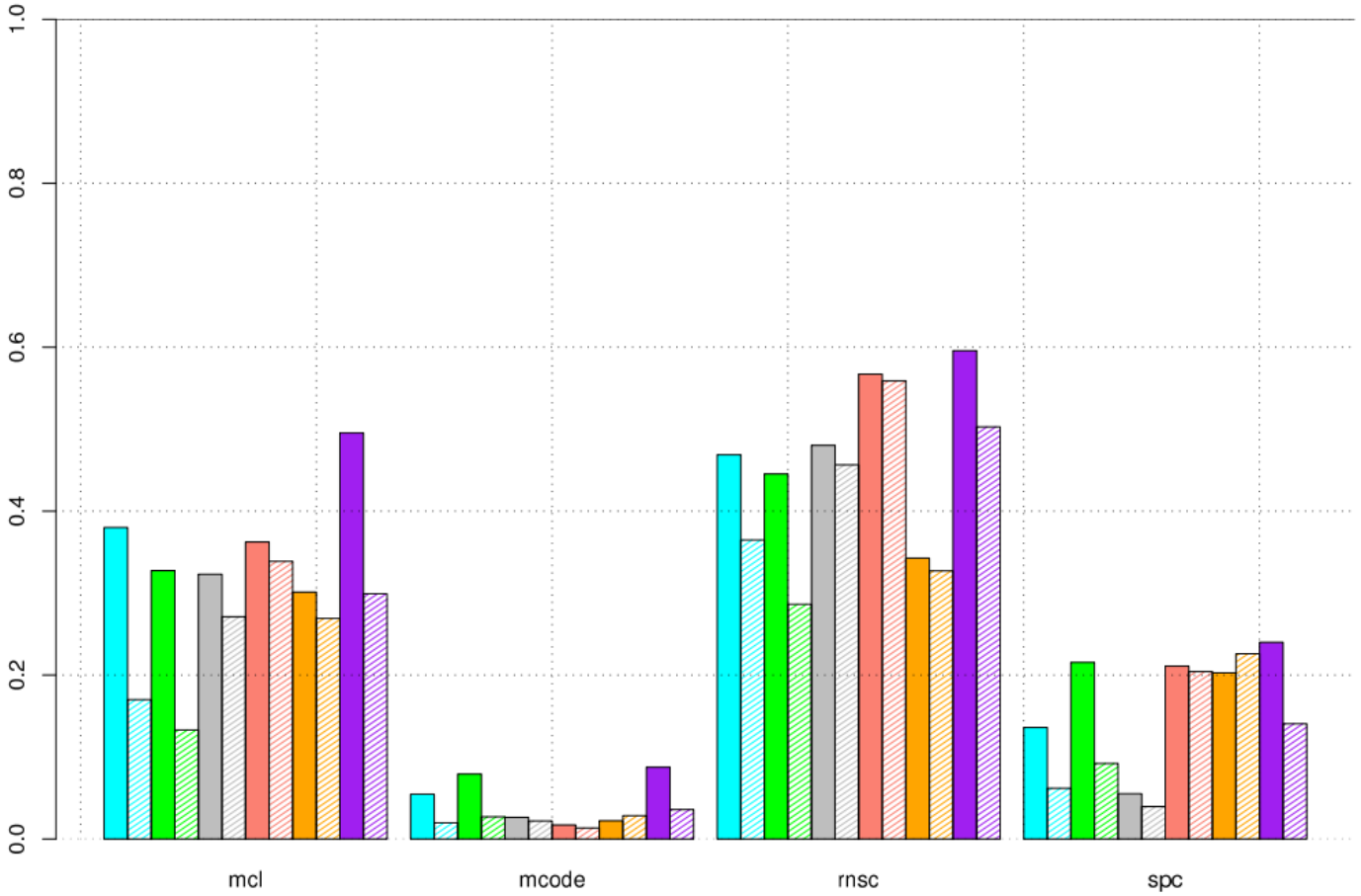


Links Genauigkeit
Rechts Separation

Blau	→ MCL
Rot	→ RNSC
Orange	→ MCODE
Grün	→ SPC

Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

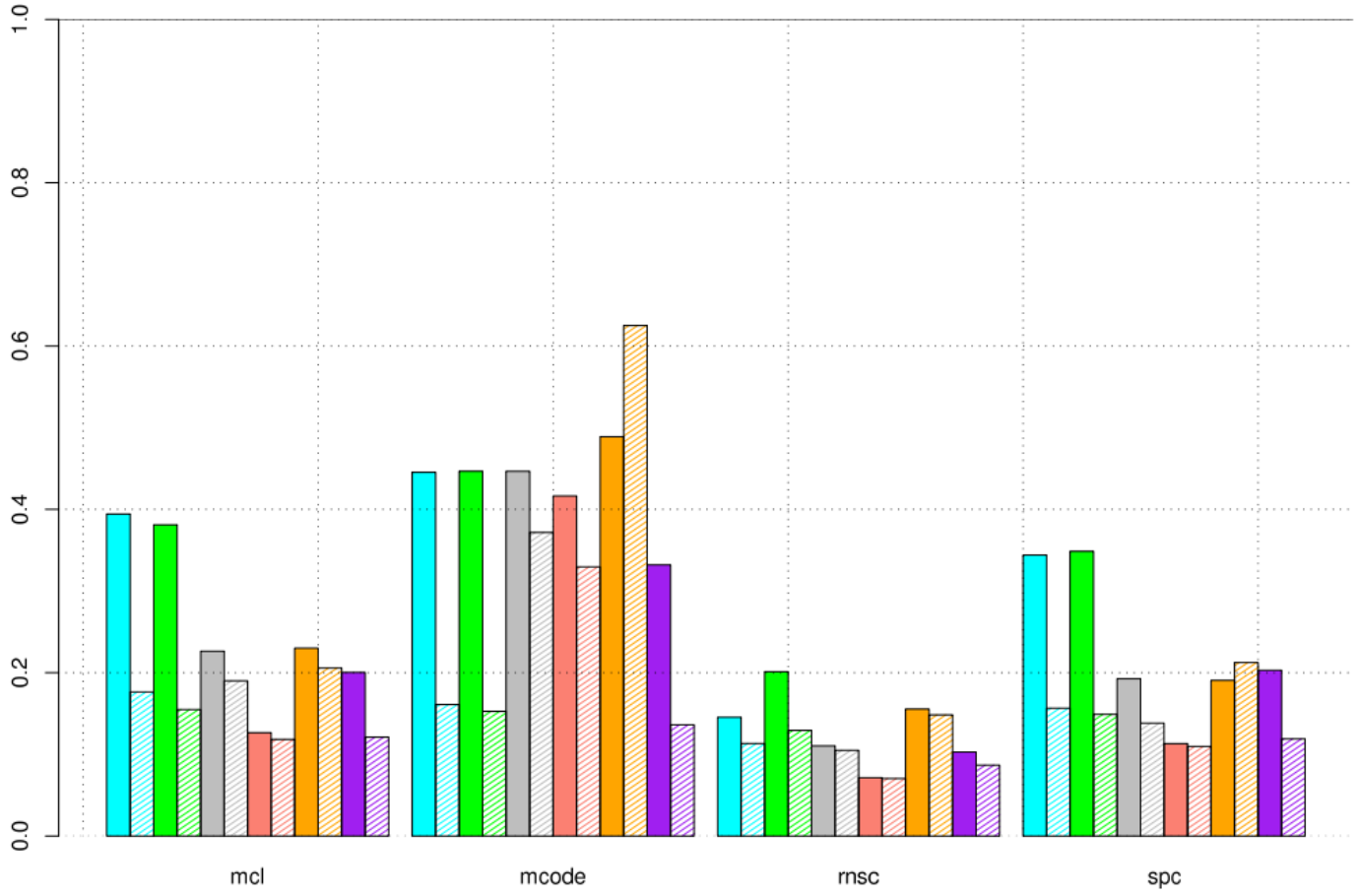
Vergleich mittels „high-throughput data sets“



Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

Cluster-
weise Separation

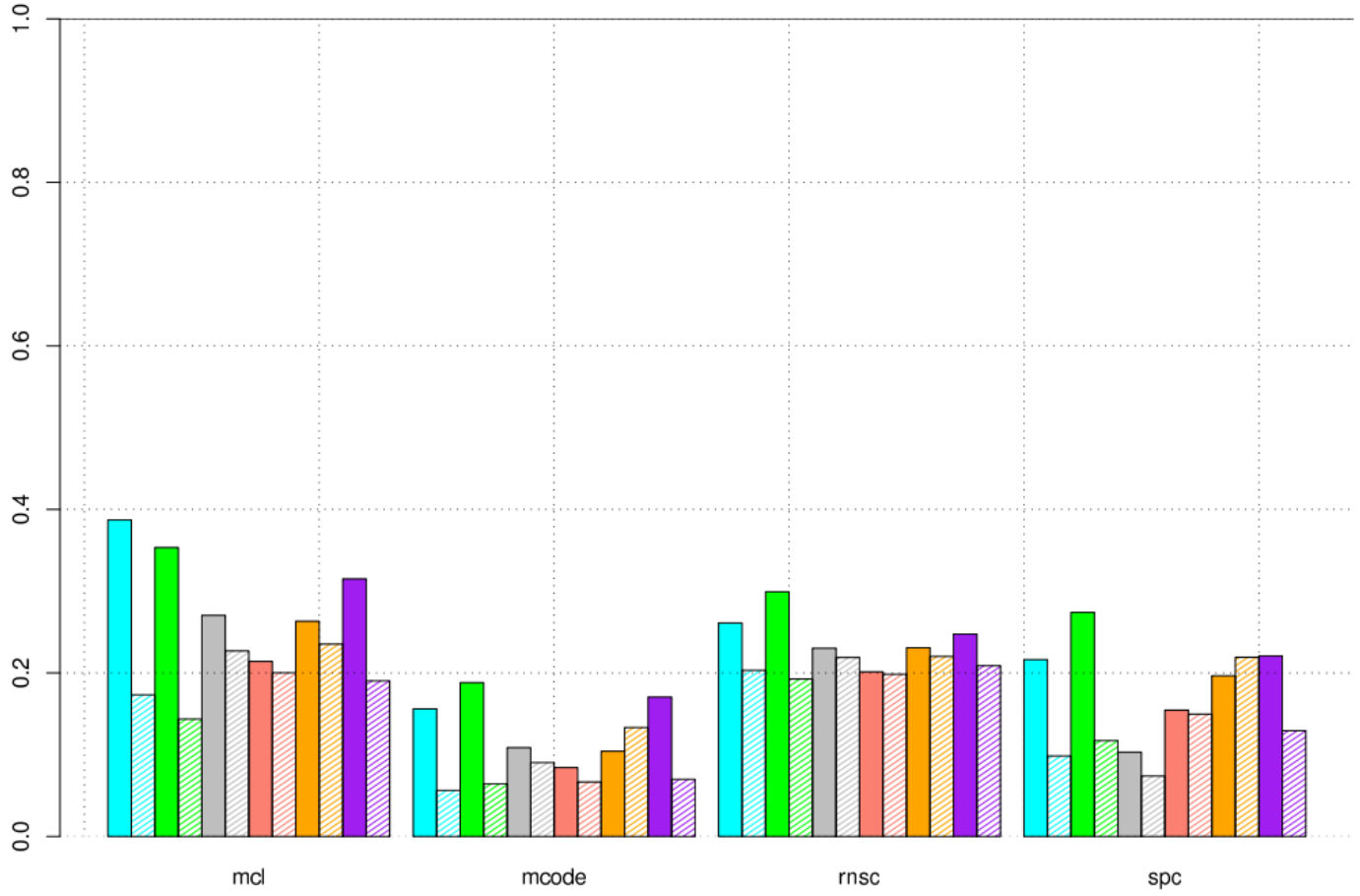
Vergleich mittels „high-throughput data sets“



Komplex-
weise Separation

Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

Vergleich mittels „high-throughput data sets“



Quelle: „Evaluation of clustering algorithms for protein-protein interaction networks“

Clustering-
weise Separation

Markov Clustering

- Fluss- basiert
- Clustering nach Wahrscheinlichkeiten
- schrittweise Transformation mittels Expansion und Inflation
- keinerlei Graphen- Vorkenntnisse erforderlich
- sehr robust gegenüber Graphänderungen
- kann mit gewichteten Kanten arbeiten
- auf vielerlei Probleme anwendbar
- Granularität kann über Parameter r bestimmt werden
- arbeitet sehr schnell
- ist gut skalierbar
- ist leicht verständlich

Quellen

MCL - a cluster algorithm for graphs

http://www.micans.org/mcl/

Evaluation of clustering algorithms for protein-protein interaction networks

http://www.biomedcentral.com/1471-2105/7/488