Axioms for graph clustering

Twan van Laarhoven and Elena Marchiori

Institute for Computing and Information Sciences
Radboud University Nijmegen, The Netherlands

27th September 2013

Outline

Outline

Introduction

Axioms for data clustering

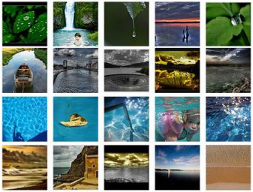
Axioms for graph clustering

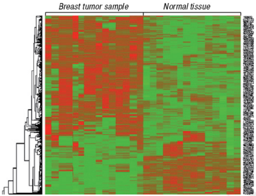
Modularity

Conclusion

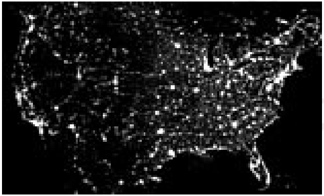
Clustering

- Image processing, medicine,



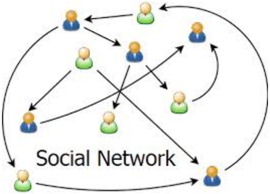


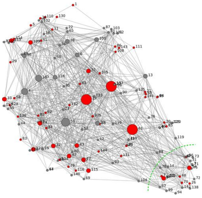
- biology, economy, ... see, e.g., UCI ML repository.



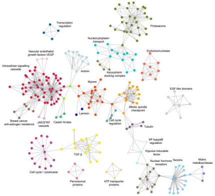
Clustering

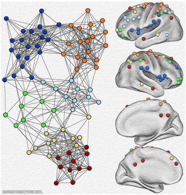
- social sciences,





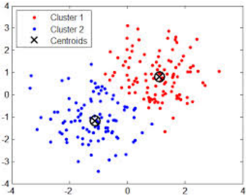
- life sciences, brain research, ... see, e.g., UCI Network Data repository.





Clustering: what is it?

- Informally: grouping objects in such a way that objects in each group are more similar to each other than to objects in other groups.



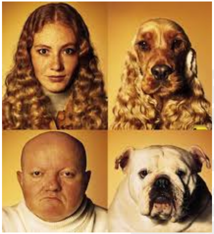
- Formally: an optimization problem. Define an objective function whose optimization yields a division of objects into (disjoint) groups. k-means clustering objective:
$\sum_{c \in C} \sum_{\vec{x} \in c} ||\vec{x} - \vec{\mu}_c||_2$, where $\vec{\mu}_c = \sum_{\vec{x} \in c} \vec{x}/|c|$.

Clustering: how to do it?

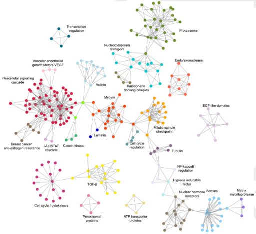
- Clustering as an optimization problem is in general NP-hard.
- Efficient heuristic and approximation algorithms are developed to find sub optimal solutions.

Clustering: data versus graphs

- Data clustering uses a *distance function* that quantifies the similarity between each pair of patterns.

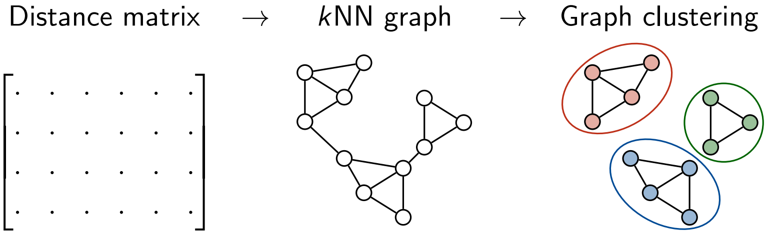


- Graph clustering uses *weighted edges* describing a relation over patterns.



From data to graph clustering

- Proximity graphs may be used to transform a data clustering problem into a graph clustering one.



Outline

Why axioms?

- There is no unique definition of clustering.
- Can we formalize our intuition of good objective functions?
- Are existing objective functions good?
- Can we design better objective functions?

Axioms for data clustering

Kleinberg' s axiomatic framework

Kleinberg proved an impossibility result concerning the axiomatization of the notion of data clustering.

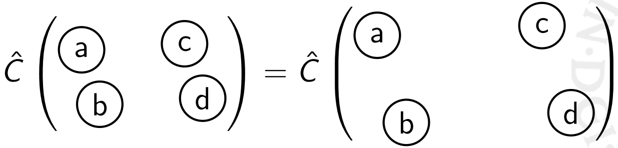
He focused on clustering functions $\hat{C} : \mathcal{D} \to \mathcal{C}$, from distance functions over a dataset S to clusterings of S, $d \mapsto C$.

Theorem (Kleinberg 2002)

There is no clustering function that is scale invariant, consistent and rich.

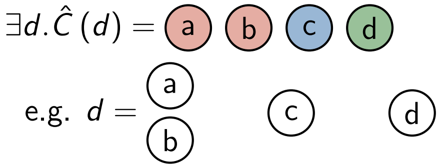
Kleinberg's axioms

- **Scale-Invariance.**
  $\forall d \in \mathcal{D}, \alpha > 0. \quad \hat{C}(d) = \hat{C}(\alpha d).$



Kleinberg's axioms

- **Richness.**
  range($\hat{C}$) is equal to the set of all partitions of S.



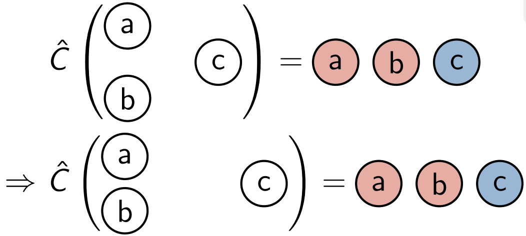
Kleinberg's axioms

- **Consistency**.

$$\forall d, d' \in \mathcal{D}. \quad \Big(\hat{C}(d) = C \text{ and } d' \text{ is a } C\text{-transformation of } d \Rightarrow \hat{C}(d') = C\Big).$$

d' is a *C-transformation of d* if $\forall i, j \in S$

  - $i \sim_C j \Rightarrow d'(i,j) \leq d(i,j)$;
  - $i \not\sim_C j \Rightarrow d'(i,j) \geq d(i,j)$.



Kleinberg's axioms

- **Scale-Invariance**.
  $\forall d \in \mathcal{D}, \alpha > 0. \quad \hat{C}(d) = \hat{C}(\alpha d).$
- **Richness**.
  range$(\hat{C})$ is equal to the set of all partitions of S.
- **Consistency**.
  $\forall d, d' \in \mathcal{D}. \quad \big(\hat{C}(d) = C$ and d' is a C-transformation of d
  $\Rightarrow \hat{C}(d') = C\big).$

  d' is a *C-transformation of d* if $\forall i, j \in S$

  - $i \sim_C j \Rightarrow d'(i,j) \leq d(i,j);$
  - $i \not\sim_C j \Rightarrow d'(i,j) \geq d(i,j).$

Kleinberg result

C' is a *refinement* of C ($C' \sqsubseteq C$) if
$\forall c' \in C' \; \exists c \in C$ s.t. $c' \subseteq c$.

$\{C_1, \ldots, C_n\} \subset \mathcal{C}$ is an *antichain* if $\forall i, j \; i \neq j \Rightarrow C_i \not\sqsubseteq C_j$.

Theorem

If $\hat{C}$ is Scale Invariant and Consistent then range($\hat{C}$) is an antichain.

Proof (sketch)

Suppose $\hat{C}$ is Consistent and Scale Invariant. Let $C_0 \sqsubseteq C_1$ in range($\hat{C}$). Construct d such that $\hat{C}(d) = C_1$. Choose α such that $d' = \alpha d$ and $\hat{C}(d') = C_0$.

Other results

Quality functions

Ackerman and Ben-David used quality functions Q instead of clustering functions. $Q : \mathcal{D} \times \mathcal{C} \to \mathbb{R}_{\geq 0}$, mapping a distance function and a clustering into a non-negative real number, $(d, C) \mapsto r$.

Theorem (Ackerman, Ben-David 2008)

There is a clustering quality function that is permutation invariant, scale invariant, monotonic and rich.

C-index $= (s - s_{min})/(s_{max} - s_{min})$, where $s = \sum_{i \sim_C j} d(i, j)$, s_{min} is the sum of the n minimal (over all pairs of patterns) distances, s_{max} is the sum of the n maximal distances, $n = |\{(i, j) \mid i \sim_C j\}|$.

To summarize

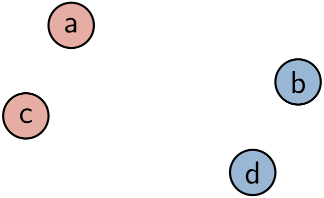
- Previous work on axioms for clustering objective functions are framed in terms of distance functions.
- Kleinberg's impossibility result is for clustering functions.
- Quality functions are more flexible and allow for axiomatization of data clustering.
- What about graph clustering? This is a different - although related - story ...

Outline

Graphs

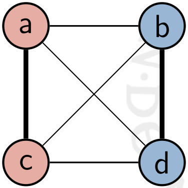
Distance functions

$d(i,j)$



Graphs

$E(i,j)$



Graphs

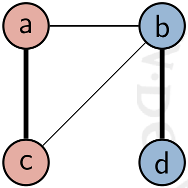
Distance functions

$d(i, j)$

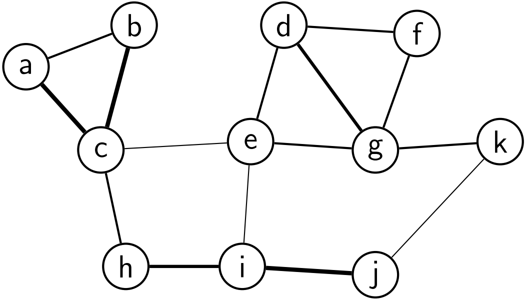
-

Graphs

$E(i, j)$



Graphs

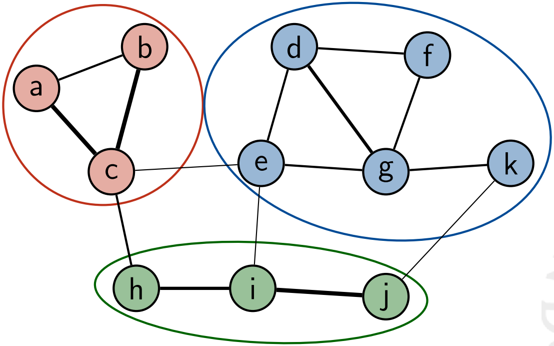


A symmetric weighted **graph** (or network) is a pair (V, E) of

- a finite set V of **nodes**, and
- a function $E : V \times V \to \mathbb{R}_{\geq 0}$ of **edge weights**,

such that $E(i, j) = E(j, i)$ for all $i, j \in V$.

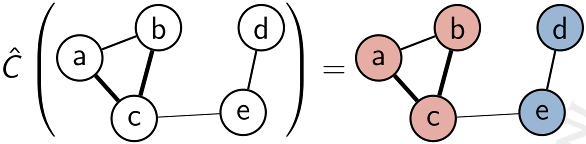
Graph clustering



A **clustering** C of a graph $G = (V, E)$ is a partition of its nodes.

Clustering: formalizations

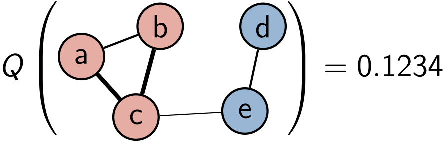
1. Clustering function
   $\hat{C}$: Graph $\rightarrow$ Clustering



2. Quality function
   Q : Graph $\times$ Clustering $\rightarrow \mathbb{R}$
3. Quality relation
   $\cdot \preceq^G \cdot \subseteq$ Clustering $\times$ Clustering

Clustering: formalizations

1. Clustering function
   $\hat{C}$: Graph $\to$ Clustering
2. Quality function
   Q : Graph $\times$ Clustering $\to \mathbb{R}$



3. Quality relation
   $\cdot \preceq^{G} \cdot \subseteq$ Clustering $\times$ Clustering

Clustering: formalizations

1. Clustering function

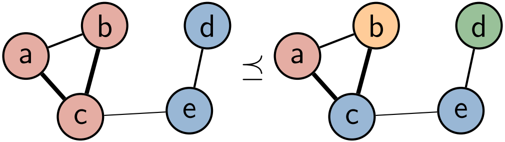
   $\hat{C}$: Graph $\to$ Clustering

2. Quality function

   Q : Graph $\times$ Clustering $\to \mathbb{R}$

3. Quality relation

   $\cdot \preceq^G \cdot \subseteq$ Clustering $\times$ Clustering



Some quality functions

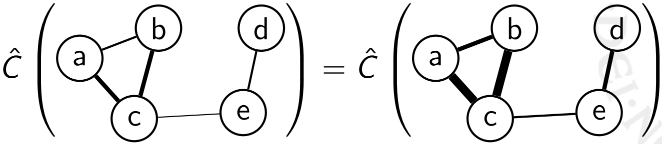
- Connected components
- Total weight of within cluster edges
$$Q(G,C) = \sum_{c \in C} w_c$$
- Modularity
$$Q(G,C) = \sum_{c \in C} \left(w_c/v_V - (v_c/v_V)^2 \right)$$
- Many more
$$Q(G,C) = \sum_{c \in C} -w_c \log(v_c/v_V)$$
...

Families of quality functions

- Connected components with threshold
- Total weight of within cluster edges with penalty
$$Q(G,C) = \sum_{c \in C} w_c - \alpha |C|$$
- Modularity
$$Q^{\gamma}_{\mathrm{RB}}(G,C) = \sum_{c \in C} \left(w_c / v_V - \gamma (v_c / v_V)^2 \right)$$
- Many more
$$Q(G,C) = \sum_{c \in C} -w_c \log(v_c / \alpha)$$
...

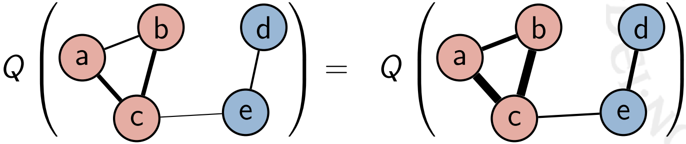
Axiom 1: Scale invariance

Intuition: The magnitude of the edge weights shouldn't matter.



Axiom 1: Scale invariance

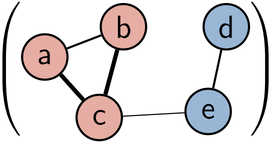
Intuition: The magnitude of the edge weights shouldn't matter.



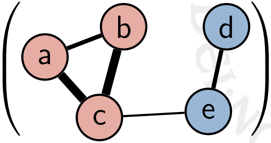
Axiom 1: Scale invariance

Intuition: The magnitude of the edge weights shouldn't matter.

Q

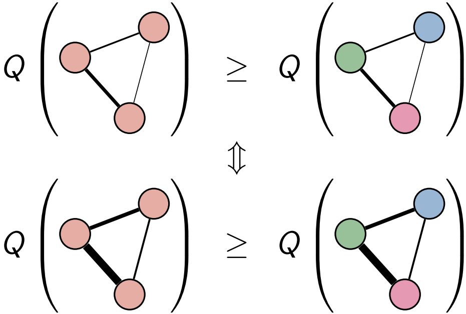


$= \alpha Q$



Axiom 1: Scale invariance

Intuition: The magnitude of the edge weights shouldn't matter.



Axiom 1: Scale invariance

Intuition: The magnitude of the edge weights shouldn't matter.

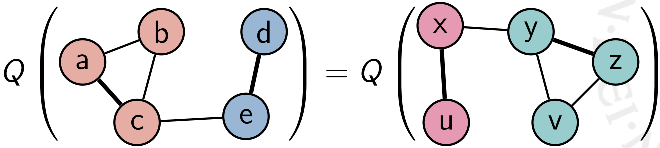
A quality function Q is **scale invariant** if

- for all graphs $G = (V, E)$,
- all constants $\alpha > 0$,

$Q(G, C_1) \geq Q(G, C_2)$ if and only if $Q(\alpha G, C_1) \geq Q(\alpha G, C_2)$.

Axiom 2: Permutation invariance

Intuition: Only the edge weights should matter.



Axiom 2: Permutation invariance

Intuition: Only the edge weights should matter.

A quality function Q is **permutation invariant** if

$$Q(G, C) = Q(f(G), f(C)).$$

for all

- graphs $G = (V, E)$ and
- all isomorphisms $f : V \to V'$,

where f is extended to graphs and clusterings in the obvious way.

Axiom 3: Richness

Intuition:

- All clusterings must be possible.

So,

- no trivial quality functions.
- no fixed number of clusters.

A quality function Q is **rich** if

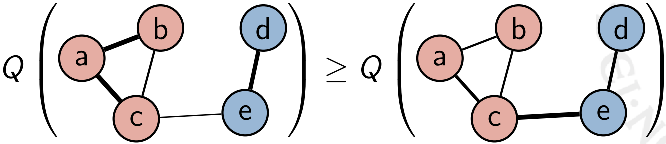
- for all sets V and
- all partitions C^* of V,

there is

- a graph $G = (V, E)$
- such that C^* is the optimal clustering of G.

Axiom 4: Monotonicity

Intuition: Adding edges inside a cluster or removing edges between clusters does not make the clustering worse.



Axiom 4: Monotonicity

Intuition: Adding edges inside a cluster or removing edges between clusters does not make the clustering worse.

Let

- $G = (V, E)$ and $G' = (V, E')$ be graphs, and
- C be a clustering of G and G'.

Then G' is a C-**consistent improvement of** G if

- $E'(i,j) \geq E(i,j)$ for all $i \sim_C j$ and
- $E'(i,j) \leq E(i,j)$ for all $i \not\sim_C j$.

Axiom 4: Monotonicity

Intuition: Adding edges inside a cluster or removing edges between clusters does not make the clustering worse.

A quality function Q is **monotonic** if

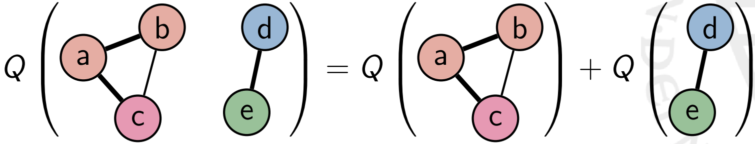
$$Q(G', C) \geq Q(G, C).$$

for all

- graphs G,
- all clusterings C of G and
- all C-consistent improvements G' of G.

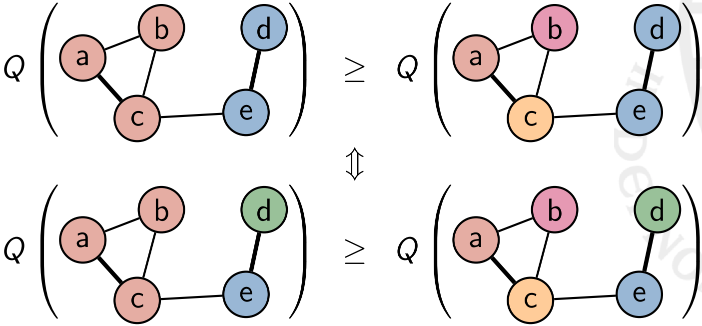
Axiom 5: Locality

Intuition: Local changes should have local effects.



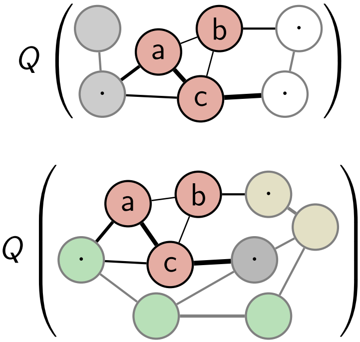
Axiom 5: Locality

Intuition: Local changes should have local effects.



Axiom 5: Locality

Intuition: Local changes should have local effects.



Axiom 5: Locality

Intuition: Local changes should have local effects.

Two graphs G_1 and G_2 **agree on the neighborhood of** $V_a \subseteq V_1 \cap V_2$ if

$$\begin{aligned} E_1(i,j) = E_2(i,j) & \text{ for all } i \in V_a,\, j \in V_1 \cap V_2, \text{ and} \\ E_1(i,j) = 0 & \text{ for all } i \in V_a,\, j \in V_1 \setminus V_2, \text{ and} \\ E_2(i,j) = 0 & \text{ for all } i \in V_a,\, j \in V_2 \setminus V_1. \end{aligned}$$

So, for nodes/clusters in V_a, all incident edges are the same.

Axiom 5: Locality

Intuition: Local changes should have local effects.

A quality function Q is **local** if

- for all graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ that agree on a set V_a and its neighborhood,
- for all clusterings C_1 of $V_1 \setminus V_a$,
  C_2 of $V_2 \setminus V_a$ and
  C_a, D_a of V_a.

if $\quad Q(G_1, C_a \cup C_1) \geq Q(G_1, D_a \cup C_1)$

then $Q(G_2, C_a \cup C_2) \geq Q(G_2, D_a \cup C_2)$.

Discontinuity is magic

Theorem

There is a graph clustering function that is scale invariant, permutation invariant, monotonic, rich and local.

$$\hat{C}_{\text{coco}}(G) = \text{the connected components of } G$$

$$Q_{\text{coco}}(G, C) = \mathbf{1}[C \text{ are the connected components of } G]$$

- Doesn't this contradict Kleinberg's theorem?
- No: edge weight $= 0 \Leftrightarrow$ distance $= \infty$.
- Connected components are unstable.

Discontinuity is magic

Theorem

There is a graph clustering function that is scale invariant, permutation invariant, monotonic, rich and local.

$$\hat{C}_{\text{coco}}(G) = \text{the connected components of } G$$

$$Q_{\text{coco}}(G, C) = \mathbf{1}[C \text{ are the connected components of } G]$$

- Doesn't this contradict Kleinberg's theorem?
- No: edge weight $= 0 \Leftrightarrow$ distance $= \infty$.
- Connected components are unstable.

Axiom 6: continuity

Intuition:

- Don't allow such unstable quality functions.
- A small change in edge weights should lead to only a small change in quality.

A quality function Q is **continuous** if

- for every $\epsilon > 0$ and
- every graph $G = (V, E)$

there exists a $\delta > 0$ such that

- for every graph $G' = (V, E')$ and
- every clustering C of G,

we have $\|E' - E\|_{\text{max}} < \delta \Rightarrow |Q(G', C) - Q(G, C)| < \epsilon$.

Outline

Modularity

Intuition:

- Balance within cluster edges against cluster volume.

$$Q_{\text{modularity}}(G, C) = \sum_{i,j \in V} \left(\frac{E(i,j)}{v_V} - \frac{v_i}{v_V}\frac{v_j}{v_V} \right) \mathbf{1}[i \sim_C j].$$

$$= \sum_{c \in C} \left(\frac{w_c}{v_V} - \left(\frac{v_c}{v_V} \right)^2 \right).$$

Where

$v_c = \sum_{i \in c} \sum_{j \in V} E(i,j)$ volume of cluster

$w_c = \sum_{i \in c} \sum_{j \in c} E(i,j)$ within cluster weight.

Properties

The obvious:

- Modularity is permutation invariant.
- Modularity is scale invariant.
- Modularity is continuous.

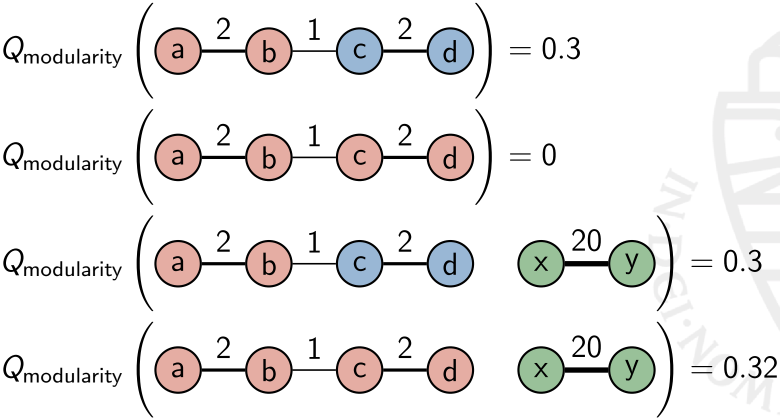
The less obvious:

- Modularity is rich.

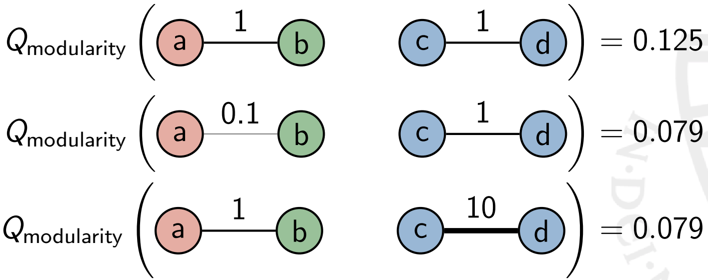
The bad:

- Modularity is *not* local.
- Modularity is *not* monotonic.

Modularity is not local



Modularity is not monotonic



Idea 1: Fix the scale

$$Q_{M\text{-fixed}}(G, C) = \sum_{c \in C} \left(\frac{w_c}{M} - \left(\frac{v_c}{M} \right)^2 \right)$$

Is it monotonic?

Take $v_c = w_c + b_c$ (within + between)

$$\frac{\partial Q_{M\text{-fixed}}(G, C)}{\partial w_c} = \frac{1}{M} - \frac{2w_c + 2b_c}{M^2}.$$

This is negative when $2v_c > M$, so not monotonic.

Idea 1: Fix the scale

$$Q_{M\text{-fixed}}(G, C) = \sum_{c \in C} \left(\frac{w_c}{M} - \left(\frac{w_c + b_c}{M} \right)^2 \right)$$

Is it monotonic?

Take $v_c = w_c + b_c$ (within + between)

$$\frac{\partial Q_{M\text{-fixed}}(G, C)}{\partial w_c} = \frac{1}{M} - \frac{2w_c + 2b_c}{M^2}.$$

This is negative when $2v_c > M$, so not monotonic.

Idea 2: Add some v_c to the denominator

$$Q_{M,\gamma}(G,C) = \sum_{c \in C} \left(\frac{w_c}{M + \gamma v_c} - \left(\frac{v_c}{M + \gamma v_c} \right)^2 \right).$$

Adaptive scale modularity is

- permutation invariant, continuous and local.
- monotonic for all $M \geq 0$ and $\gamma \geq 2$.
- rich for all $M \geq 0$ and $\gamma \geq 1$.
- scale invariant for $M = 0$.

Idea 2: Add some v_c to the denominator

$$Q_{M,\gamma}(G,C) = \sum_{c\in C}\left(\frac{w_c}{M+\gamma v_c} - \left(\frac{v_c}{M+\gamma v_c}\right)^2\right).$$

Adaptive scale modularity is

- permutation invariant, continuous and local.
- monotonic for all $M \geq 0$ and $\gamma \geq 2$.
- rich for all $M \geq 0$ and $\gamma \geq 1$.
- scale invariant for $M = 0$.

Proof of monotonicity

Take partial derivatives ($v_c = w_c + b_c$)

$$Q_{M,\gamma}(G,C) = \sum_{c\in C}\left(\frac{w_c}{M+\gamma(w_c+b_c)} - \left(\frac{w_c+b_c}{M+\gamma(w_c+b_c)}\right)^2\right).$$

$$\frac{\partial Q_{M,\gamma}(G,C)}{\partial w_c} = \frac{M^2+(\gamma-2)Mw_c+(2\gamma-2)Mb_c+\gamma^2 v_c b_c}{(M+\gamma v_c)^3}.$$

$$\frac{\partial Q_{M,\gamma}(G,C)}{\partial b_c} = -\frac{2Mv_c}{(M+\gamma v_c)^3} - \frac{\gamma w_c}{(M+\gamma v_c)^2} \le 0.$$

When $\gamma \ge 2$, Q is a monotonic increasing function of w_c and decreasing function of b_c for all c, so the quality function is monotonic. □

Proof sketch of richness

- Given a clustering C^* take G to be the clique graph of C^*.
- Pick edge weight large enough ($k > 2|V|^3 M$), then the effect of M becomes insignificant.

$$Q(G, D) \approx \sum_{c \in C} \left(w_d - \frac{v_d^2}{\gamma v_d} \right).$$

- There are at most $|C^*|$ terms in the sum that are $> \epsilon$ (where ϵ depends on k and M)
- The term for $c \in C$ is maximal if $c = \bigcup D, D \subseteq C^*$.

The **clique graph** with edge weight k of a partition C of V is (V, E) where $E(i, j) = k \cdot \mathbf{1}[i \sim_C j]$.

Related quality functions

- When $\gamma = 0$, we get fixed scale modularity. Equivalent to other modularity variants.
- When $\gamma = 0$ and $M = v_V$, we get modularity.
- When $M = 0$ we get

$$Q_{0,\gamma}(G, C) \propto \sum_{c \in C} \left(\frac{w_c}{v_c} - \frac{1}{\gamma} \right),$$

  i.e. normalized cut.
- When $M \to \infty$ we get

$$Q_{\infty,\gamma}(G, C) \propto \sum_{c \in C} w_c,$$

  i.e. unnormalized cut.

Outline

Summary

- Graph and data clustering are related, yet different, notions.
- 6 axioms for graph clustering quality functions.
- Graph setting allows for locality.
- Modularity is not monotonic.
- Adaptive scale modularity satisfies all 6 axioms.
- Generalizes both modularity and normalized cut.
- Two parameters to control size of clusters.

Open problems

- Applications of adaptive scale modularity to real life problems.
- Overlapping clusters.
- Directed graphs.
- How to use axioms for developing better algorithms for clustering.

Thank you for your attention.

Axioms for graph clustering

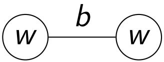
Twan van Laarhoven and Elena Marchiori

Institute for Computing and Information Sciences
Radboud University Nijmegen, The Netherlands

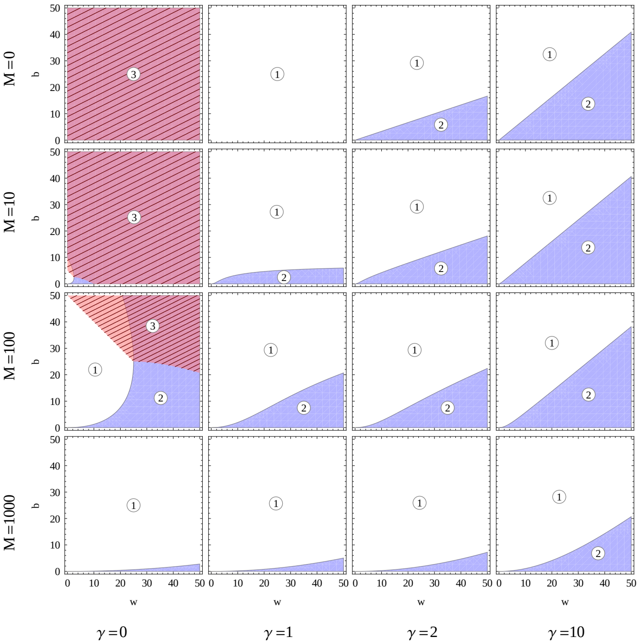
27th September 2013

Extra slides

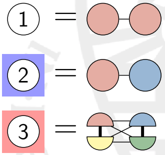
Adaptive Scale Modularity behavior

Take a simple graph: 

- Two cliques each with w within weight
- Connected by edges with total weight b.
- Total volume $2w + 2b$.
- What is the behavior of adaptive scale modularity?



Legend:



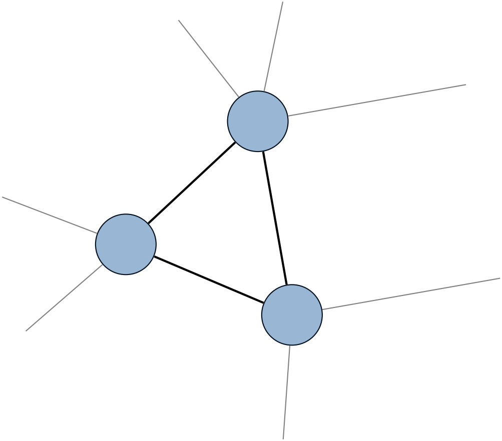
Clustering by optimization

- Graph clustering is NP hard.
- Top down:
  find best cut and repeat
- Bottom up:
  group nodes together
- Simulated annealing

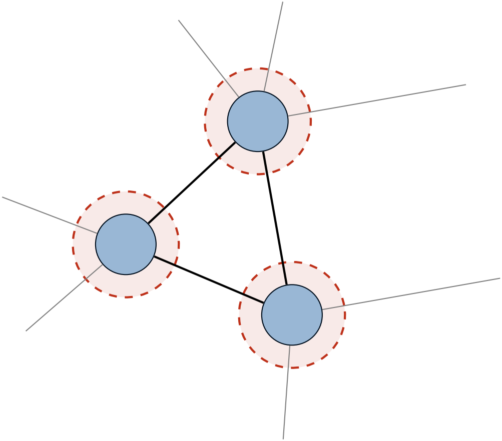
Louvain method

- V.D. Blondel, JL. Guillaume, R. Lambiotte, E. Lefebvre
  Fast unfolding of communities in large networks
  J. Stat. Mech. 2008
- Best graph clustering method in surveys.
- Method:
  1. Move nodes into neighboring clusters to improve quality.
  2. Repeat until local maximum.
  3. Now cluster the clusters.

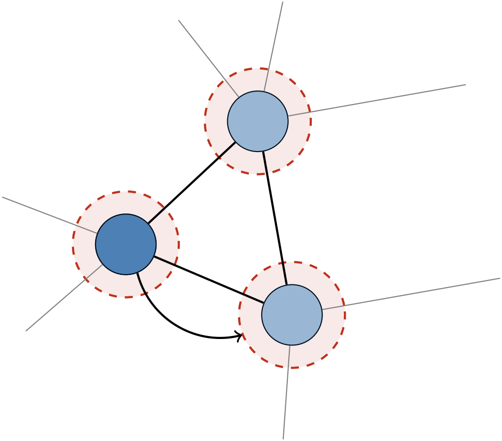
Louvain method (example)



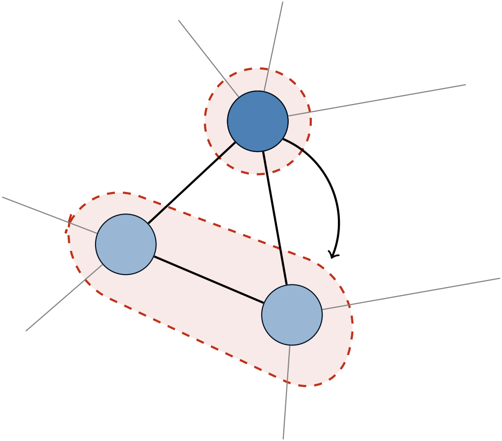
Louvain method (example)



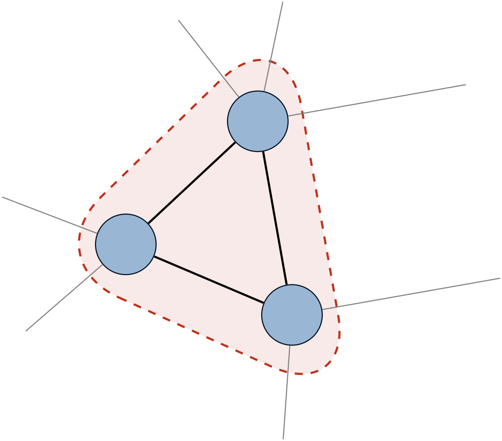
Louvain method (example)



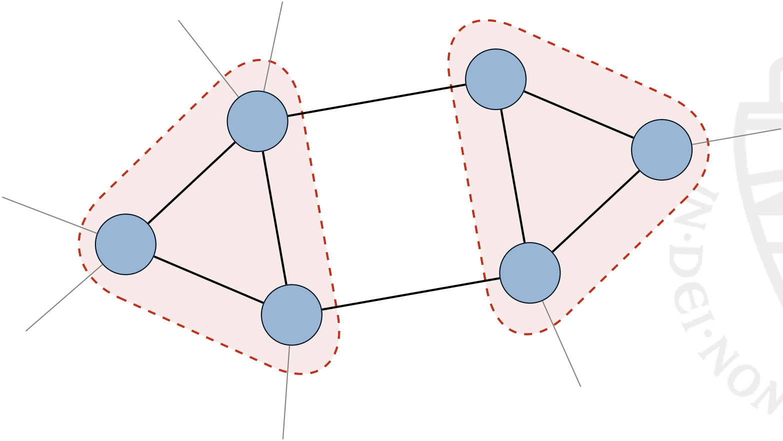
Louvain method (example)



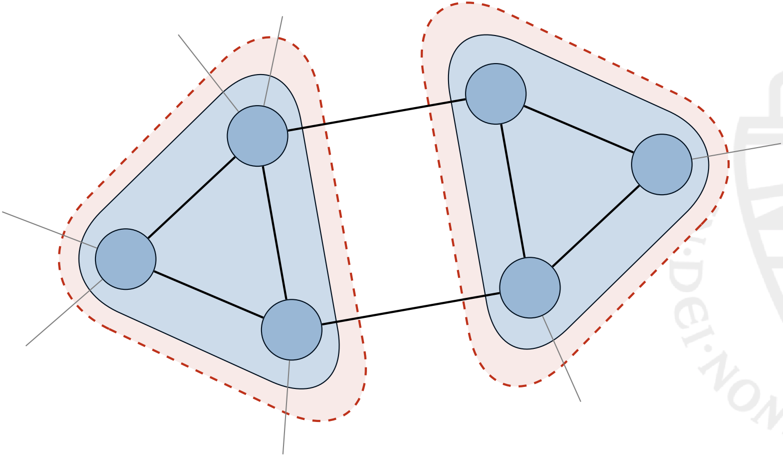
Louvain method (example)



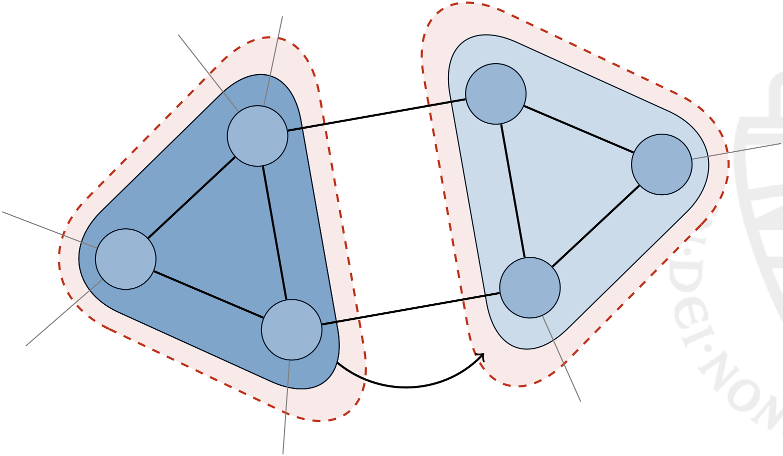
Louvain method (example)



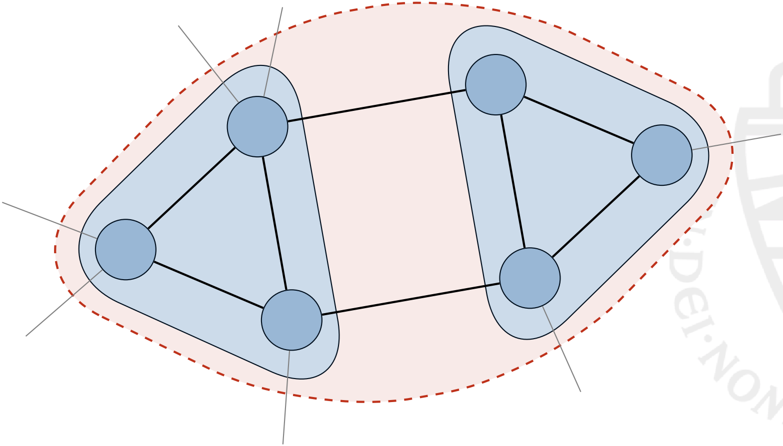
Louvain method (example)



Louvain method (example)



Louvain method (example)



Louvain method (example)