Beyond Spectral Clustering: A Comparative Study of Community Detection for Document Clustering

Kostadin Cvejoski

Fraunhofer Institute for
Intelligent Analysis and Information Systems IAIS

Methodology

- embed the documents in some vector space
- define graph over the vector space
- applay the Overlap Hierarchical Algorithm

Basic Notations

Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

- $\mathcal{V}$ - *nodes*; $\mathcal{V} = \{v_i\}_{i=1}^{n}$, where $n = |\mathcal{V}|$,
- $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$ - *edges*

Subgraph $\mathcal{C} = (\mathcal{V}(\mathcal{C}), \mathcal{E}(\mathcal{C}))$ of a graph $\mathcal{G}$:

- $\mathcal{V}(\mathcal{C}) \subseteq \mathcal{V}$ - *nodes*; $n_c = |\mathcal{C}|$,
- $\mathcal{E}(\mathcal{C}) = \{(v_i, v_j) : v_i, v_j \in \mathcal{V}(\mathcal{C}) \wedge (v_i, v_j) \in \mathcal{E}\}$ - *edges*

Partition $\mathcal{P}$ of a graph $\mathcal{G}$:

$$\mathcal{P} \equiv \{\mathcal{C}_1, ..., \mathcal{C}_M\} \text{ such that } \bigcup_{i=1}^{M} \mathcal{V}(\mathcal{C}_i) = \mathcal{V}(\mathcal{G})$$

Overlap Hierachical Algorithm (OH) - Notations

- **internal degree**$(k_v^{int}(\mathcal{C}))$ of $v \in \mathcal{C}$ # of edges with nodes in $\mathcal{C}$,
- **external degree**$(k_v^{ext}(\mathcal{C}))$ of $v \in \mathcal{C}$ # of edges with nodes not in $\mathcal{C}$,
- **internal(external)** degree of subgraph $\mathcal{C}$:

$$k_{\mathcal{C}}^{int(ext)} = \sum_{v \in \mathcal{C}} k_v^{int(ext)}(\mathcal{C})$$

OH - Objective Function

- **fitness** $f_{\mathcal{C}}$ of a subgraph $\mathcal{C}$

$$f_{\mathcal{C}} = \frac{k_{\mathcal{C}}^{int}}{(k_{\mathcal{C}}^{int} + k_{\mathcal{C}}^{ext})^{\alpha}}$$

- **node fitness** $f_{\mathcal{C}}^{v}$ with respect to a subgraph $\mathcal{C}$:

$$f_{\mathcal{C}}^{v} = f_{\mathcal{C}+\{v\}}^{v} - f_{\mathcal{C}-\{v\}}^{v}$$

OH-Algorithm

Algorithm 1 Natural community of a node v_s

Require: Graph $\mathcal{G} \equiv (\mathcal{E}, \mathcal{V})$, v (seed node)

1: $\mathcal{C} \leftarrow \{v\}$
2: **while** there exists a node $v \in \Gamma(\mathcal{C})$ such that $f_{\mathcal{C}}^{v} > 0$ **do**
3: $\quad u = \underset{\hat{u} \in \Gamma(\mathcal{C})}{\operatorname{argmax}} f(\hat{u})$
4: $\quad \mathcal{C} \leftarrow \mathcal{C} \cup \{u\}$
5: $\quad$ **while** there exist $\tilde{u} \in \mathcal{C}$ such that $f_{\mathcal{C}}^{\tilde{u}} < 0$ **do**
6: $\quad\quad \mathcal{C} \leftarrow \mathcal{C} \setminus \{\tilde{u}\}$
7: $\quad$ **end while**
8: **end while**
9: **return** $\mathcal{C}$

$\Gamma(\mathcal{C}) = \{u | u \in \mathcal{V}$ and $u \notin \mathcal{C}$ and $(u, v) \in \mathcal{E}\}$ - neighbors of community $\mathcal{C}$

Performance Improvement of OH Algorithm

- inclusion new node in the subgraph

$$\begin{aligned} k_{\mathcal{C}}^{ext} &\leftarrow k_{\mathcal{C}}^{ext} - k_g^{int} + k_g^{ext} \\ k_{\mathcal{C}}^{int} &\leftarrow 2((k_{\mathcal{C}}^{int}/2) + k_g^{int}) \end{aligned}$$

- removing node from the subgraph

$$\begin{aligned} k_{\mathcal{C}}^{ext} &\leftarrow k_{\mathcal{C}}^{ext} + k_t^{int} - k_t^{ext} \\ k_{\mathcal{C}}^{int} &\leftarrow 2((k_{\mathcal{C}}^{int}/2) - k_t^{int}) \end{aligned}$$

Equivalence to Ncut

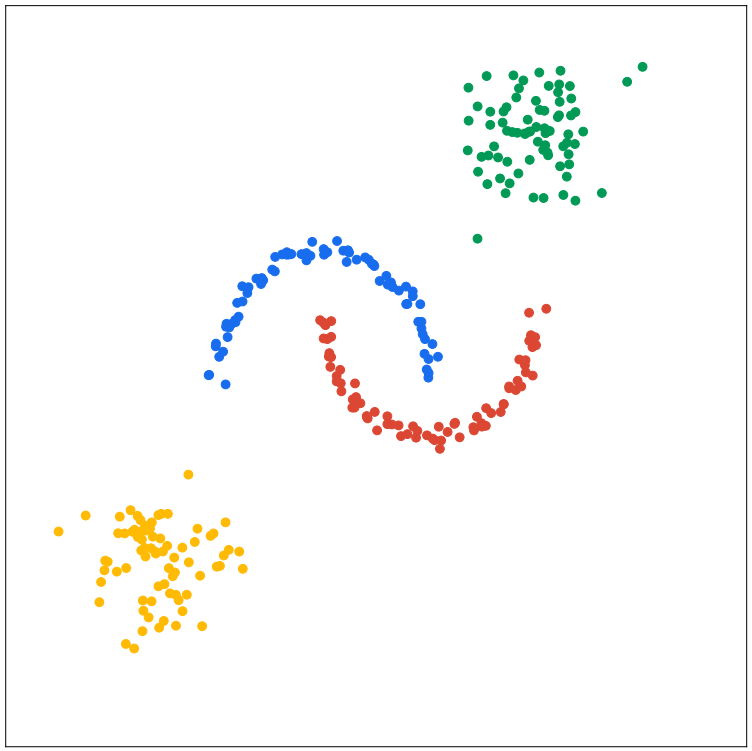
$$f(\mathcal{C}_i) = \frac{\mathsf{vol}(\mathcal{C}_i) - W(\mathcal{C}_i, \bar{\mathcal{C}}_i)}{\mathsf{vol}(\mathcal{C}_i)^{\alpha}} = \frac{1}{\mathsf{vol}(\mathcal{C}_i)^{\alpha-1}} \Big[1 - \frac{W(\mathcal{C}_i, \bar{\mathcal{C}}_i)}{\mathsf{vol}(\mathcal{C}_i)}\Big]$$

$$F_{\mathcal{P}} = \sum_i f(\mathcal{C}_i) = \sum_i \Big[1 - \frac{W(\mathcal{C}_i, \bar{\mathcal{C}}_i)}{\mathsf{vol}(\mathcal{C}_i)}\Big]$$

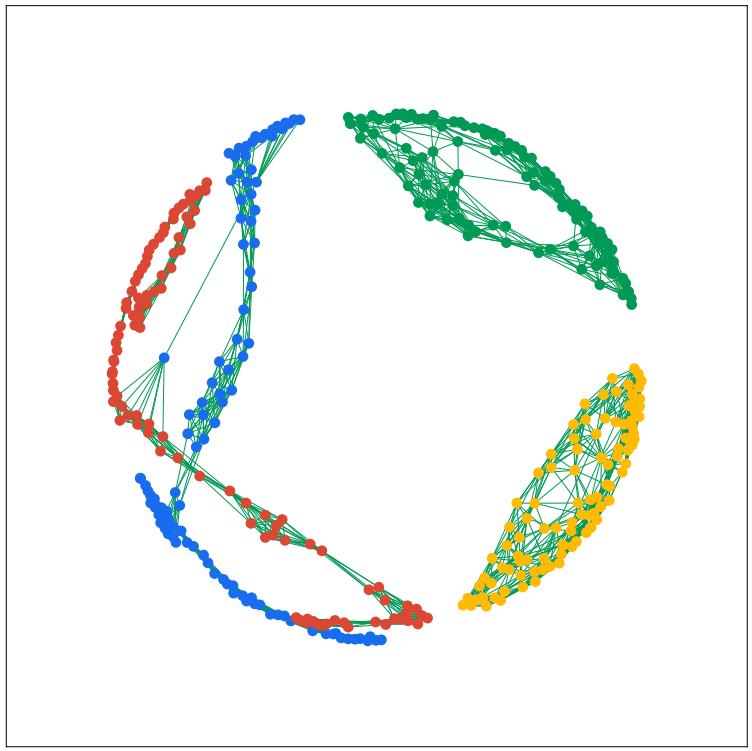
$$\min\{-F_{\mathcal{P}}\} = \min[2\mathsf{cut}(\mathcal{P}) - M] = \min\{\mathsf{cut}(\mathcal{P})\}$$

$$\mathsf{cut}(\mathcal{P}) = \sum_i \frac{W(\mathcal{C}_i, \bar{\mathcal{C}}_i)}{\mathsf{vol}(\mathcal{C}_i)}$$

Synthetic Data

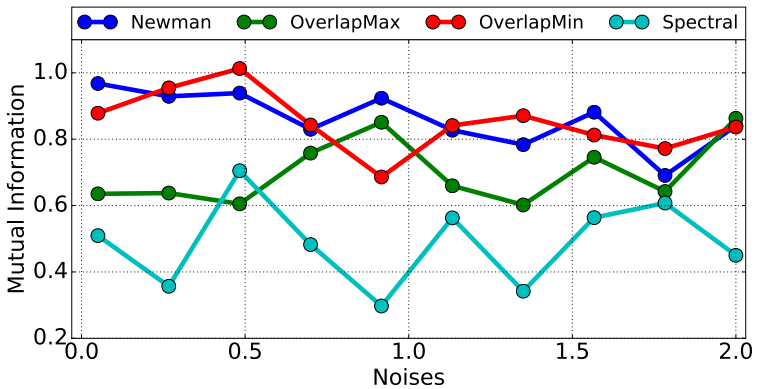


(a) spatial data

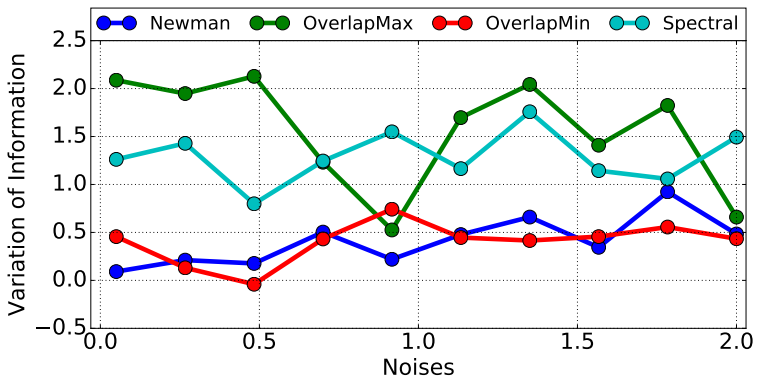


(b) associated k-NN graph

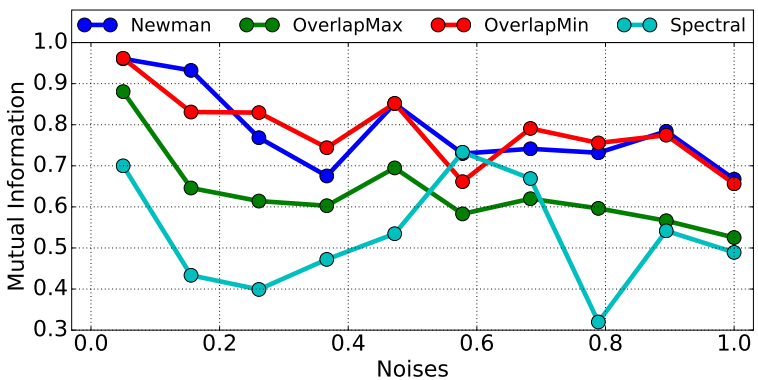
Synthetic Data



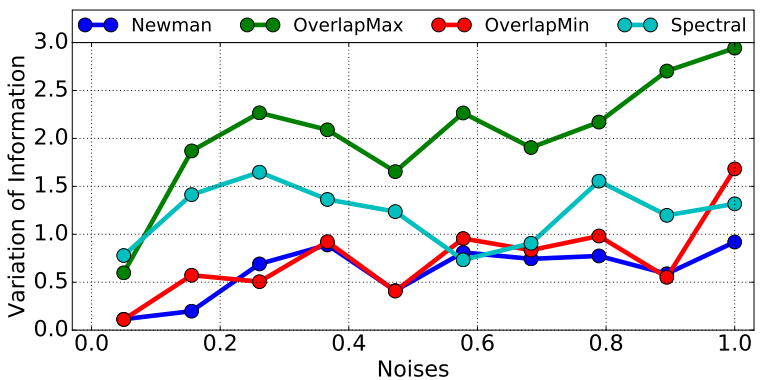
(a) Gaussian Cluster Dispersion



(b) Gaussian Cluster Dispersion



(c) Moon Cluster Dispersion



(d) Moon Cluster Dispersion

Clustering of News Articles

Com. 1: Barack Obama	
“The Obama administration rules out raising medicaid eligibility age cat food stocks plummet", “This racist snowbilly was almost our vice president palin tweets that Obama is using shuck jive", “Photos and quotes from president Barack Obamas speech in cleveland ohio", "President Obama weighs in on Michigan’s right to work for less attack on unions"	
Subcom. 1.1: Republicans vs Obama	**Subcom. 1.2: Obama Social Events**
“President Obama comes out of the gate powerfully against childish republicans mccain graham", “This racist snowbilly was almost our vice president palin tweets that Obama is using shuck jive", “The republican party is done at the national level"	"First official Obama rally of 2012 in columbus ohio", “Exclusive interactive panoramic images from the last night of the democratic national convention", “There is nothing more powerful than ordinary citizens coming together for a just cause", “Michelle Obamas 2012 dnc speech with photos and transcript"

Clustering of News Articles

Com. 2: Obamacare	
"Obamacare signups Friday 970k or 1 3m private enrollments 3 3m total", "It's not your freedom youre worried about its theirs", "Debunking the republican lie that health insurance costs have skyrocketed under Obama", "Obamacare enrollment picks up as more Americans are getting covered", "Lowering budget with health insurance", "Republican Mike Shirkey working hard to solve an Obamacare problem that literally does not exist"	
Subcom. 2.1: Republicans	**Subcom. 2.2: Insurance**
"The president almost becomes his own anger translator over GOP Obamacare sabotage", "Republicans have now entered the hypocrisy zone on the Obamacare", "GOP Obamacare replacement shows conservatives have lost the health care battle"	"Funding for Michigan health centers will help the uninsured find coverage', "Obamacare day one my hideous experience", "More on GOP bill to require insurers to tell you how Obamacare is raising your rates the house republicans respond", "RNC fundraising email claims health insurance costs under Obama have gone up eight times higher than they have"

Conclusion

- introduce new method of document clustering
- improve the performance of the OH algorithm
- equivalence to Ncut

Conclusion

- introduce new method of document clustering
- improve the performance of the OH algorithm
- equivalence to Ncut

Thank You