Overlapping community detection in labeled graphs

Esther Galbrun, Aristides Gionis and Nikolaj Tatti
galbrun@cs.bu.edu, Department of Computer Science, Boston University, US-MA
aristides.gionis@aalto.fi and nikolaj.tatti@aalto.fi, Helsinki Institute for Information Technology, Aalto University, Finland

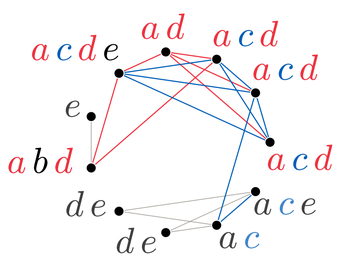
Looking for communities that are not only **dense subgraphs**,
but that also admit **compact descriptions** in terms of vertex labels.

DEFINITIONS

The *density* of a subgraph H=(U,F) is d(H) =2|F|/|U|.

We consider the *conjunctive predicate* over labels and vertices p(S) = {v ∈ V | S ⊆ (v)}.

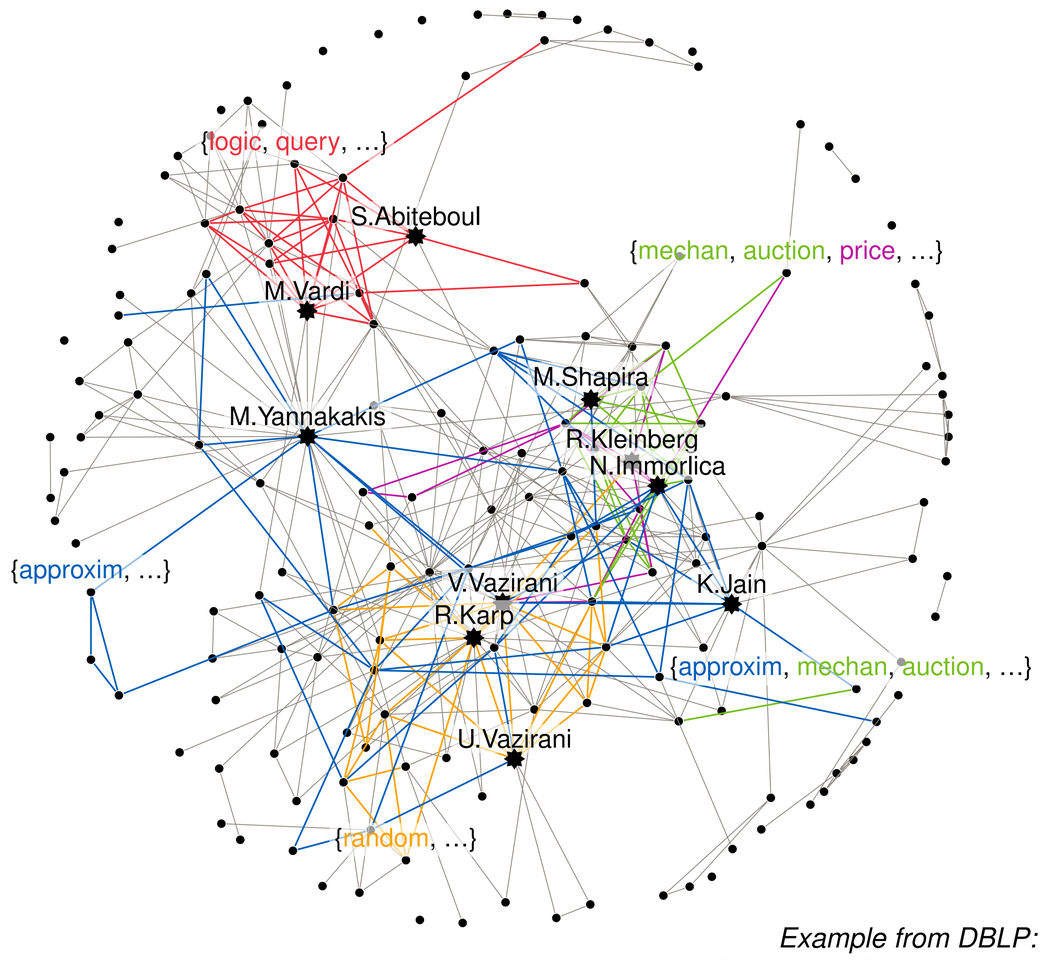
H = (U, F) is a **label-induced-subgraph** if
(i) U = p(S),
(ii) F ⊆ E (p(S))



PROBLEM STATEMENT

Let G = (V, E, l) be a multi-labeled graph,
let p be a 0-1 predicate over graph vertices and label sets,
and let k be a budget parameter.

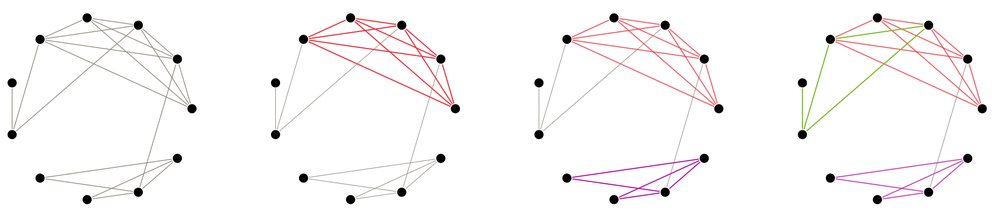
The goal is to find k sets of labels S_1, ... , S_k, and
k **disjoint** sets of edges F_1, ... , F_k, so that
(i) each $H_i = (p(S_i) , F_i)$ is a label-induced-subgraph, and
(ii) the sum of densities over all the subgraphs H_i is maximized.



Example from DBLP:
Communities within the co-authors of Christos H. Papadimitriou.

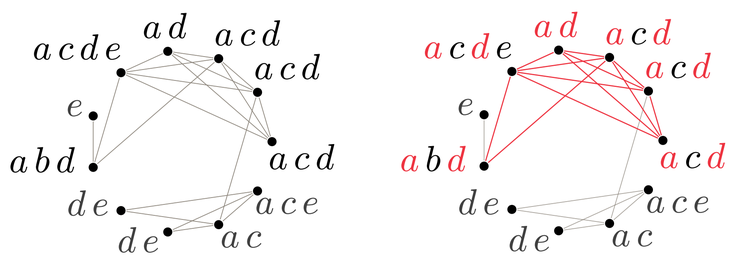
UNLABELED GRAPHS, *GMC* AND *DS*

Our algorithm couples two approximation algorithms:
(i) the Generalized Maximum-Coverage problem, by Cohen and Katzir (2008).
A variant of the max k-cover problem where elements elements have different rewards for each bin.
(ii) the Densest-Subgraph problem, by Charikar (2000).
Finding a subgraph maximizing the density.



Dense: Greedily peeling-off vertices
For i = 1 to k
(i) Remove iteratively the vertex with smallest degree,
reintroducing edges if the global score improves.
(ii) Pick among obtained graphs the one having the highest degree.
(iii) Tentatively take out edges assigned to the selected subgraph.

THREE GREEDY VARIANTS



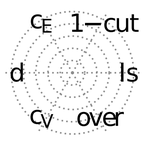
LDense: Using the labels to guide the peeling-off process.

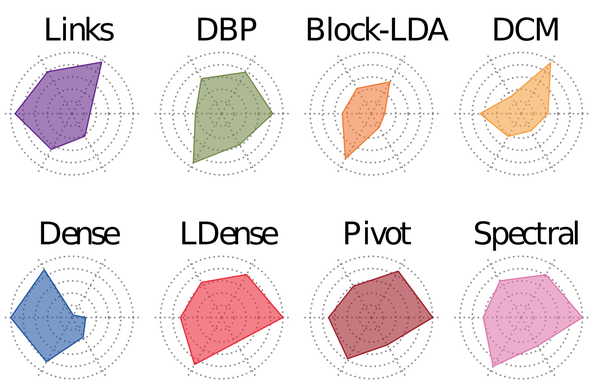
Spectral: Considering only contiguous sets of labels when ordered by the Fiedler vector of their similarity matrix.

Pivot: The local neighborhood of each vertex, "pivoted subgraphs", provide initial candidates, refined by enforcing labels predicate.

EXPERIMENTS

Comparing aspects of the communities obtained by different methods.





REFERENCES

[Links] Ahn, Bagrow and Lehmann (2010) Link communities reveal multiscale complexity in networks. Nature.
[DBP] Miettinen, Mielikäinen, Gionis, Das and Mannila (2008) *The discrete basis problem*. TKDE.
[Block-LDA] Balasubramanyan and Cohen (2011) *Block-LDA: Jointly modeling entity-annotated text and entity-entity links.* In SDM'11.
[DCM] Pool, Bonchi and van Leeuwen (2014) *Description-driven community detection.* ACM TIST.

BOSTON UNIVERSITY

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Aalto University
School of Science