KDD 2012 BEIJING
THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
Beijing, China August 12-16, 2012

Chromatic Correlation Clustering

Yahoo!

Francesco Bonchi Aristides Gionis Francesco Gullo Antti Ukkonen

Yahoo! Research – Barcelona
{bonchi,gionis,gullo,aukkonen}@yahoo-inc.com

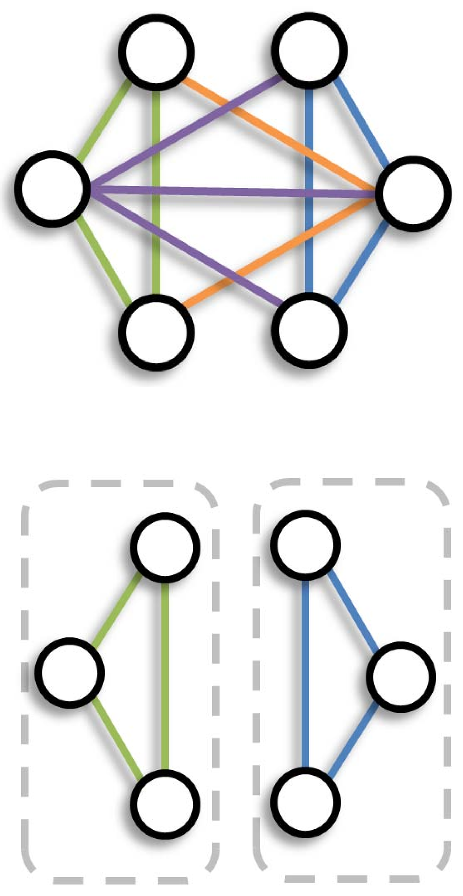
➢ We study a novel clustering problem in which the pairwise relations between objects are **categorical**. This problem can be viewed as clustering the vertices of a graph whose edges have different types (**colors**).

➢ **Applications**: social networks, protein-to-protein interaction networks, bibliographic networks, and more.

➢ We define an **objective function** to partition the graph so that the edges in each cluster have, as much as possible, the same color.

➢ The problem is **NP-hard**. We propose an ***approximation algorithm*** with provable guarantee, as well as two practical ***heuristic algorithms***.

➢ Experimental evidence on **synthetic** and **real** datasets show that our algorithms outperform a baseline algorithm both in the task of reconstructing a ground-truth clustering and in terms of objective function value.



Problem definition:

from CORRELATION CLUSTERING…

Given a set of objects V and a pairwise similarity function $\mathrm{sim}: V \times V \to [0,1]$, find a clustering $\mathcal{C}: V \to \mathbb{N}$ that minimizes the cost

$$\mathrm{cost}(\mathcal{C}) = \sum_{\substack{(x,y)\in V\times V \\ \mathcal{C}(x)=\mathcal{C}(y)}} (1-\mathrm{sim}(x,y)) + \sum_{\substack{(x,y)\in V\times V \\ \mathcal{C}(x)\neq\mathcal{C}(y)}} \mathrm{sim}(x,y).$$

… to CHROMATIC CORRELATION CLUSTERING

Given a set V of objects, a set L of labels, a special label l_0, and a pairwise labeling function $\ell: V \times V \to L \cup \{l_0\}$, find a clustering $\mathcal{C}: V \to \mathbb{N}$ and a cluster labeling function $c\ell: \mathcal{C}[V] \to L$ so to minimize the cost

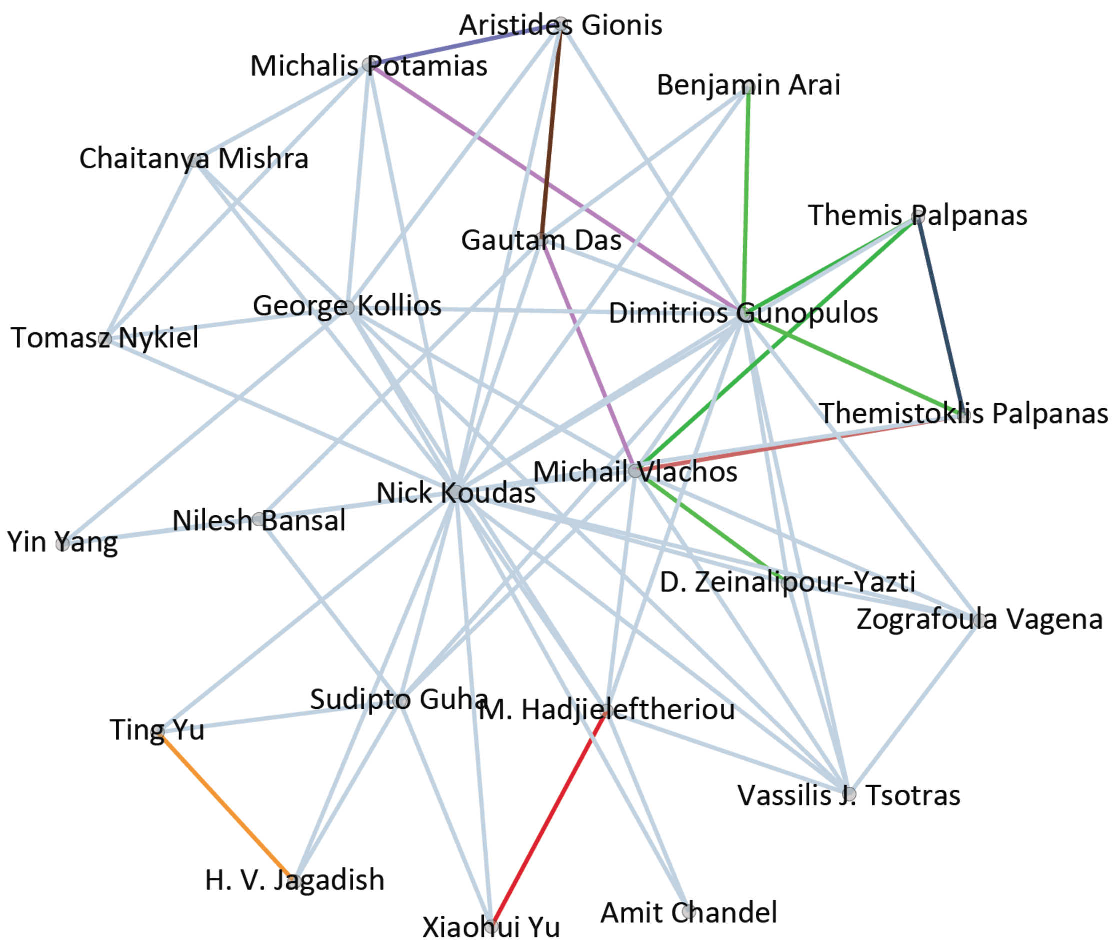
$$\mathrm{cost}(\mathcal{C}, c\ell) = \sum_{\substack{(x,y)\in V\times V, \\ \mathcal{C}(x)=\mathcal{C}(y)}} (1-\mathrm{I}[\ell(x,y) = c\ell(\mathcal{C}(x))]) + \sum_{\substack{(x,y)\in V\times V, \\ \mathcal{C}(x)\neq\mathcal{C}(y)}} \mathrm{I}[\ell(x,y) \neq l_0].$$

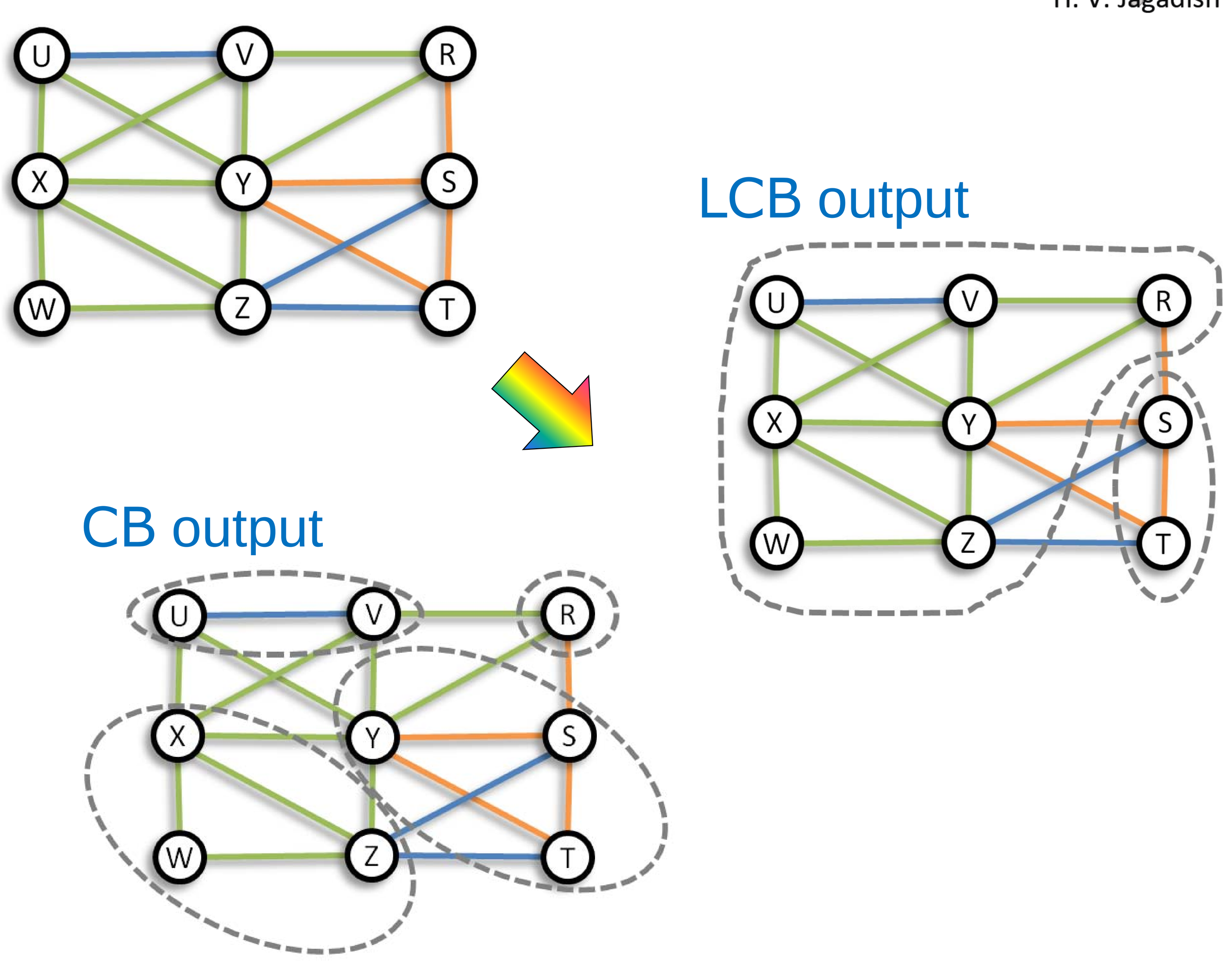
Solutions:

➢ **Randomized approximation algorithm** (CB) with approximation guarantee proportional to the maximum degree in the graph:

$$r(G) \leq 6\,(2D_{max} - 1)$$

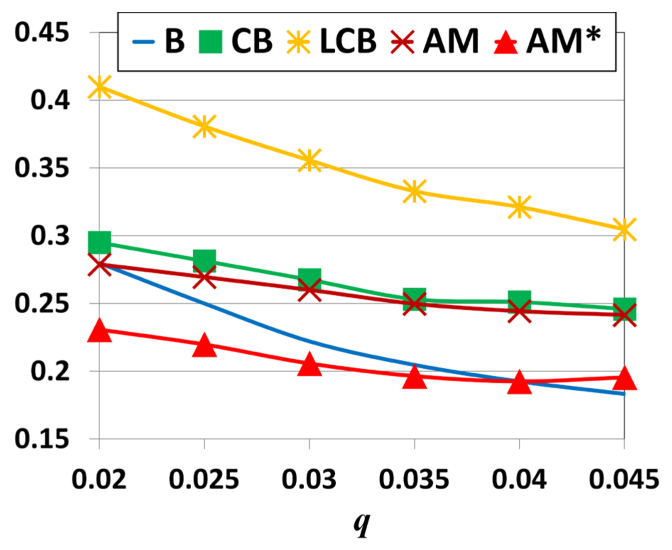
➢ Lazy CB (LCB) **algorithm**. The random choices are "guided" by heuristic considerations

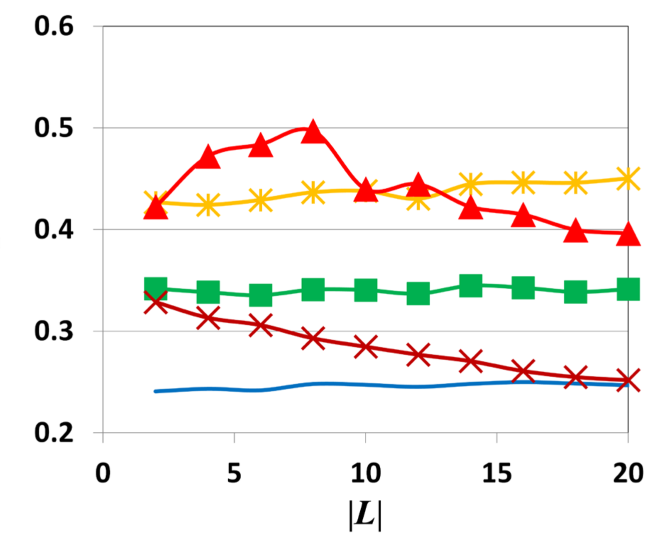


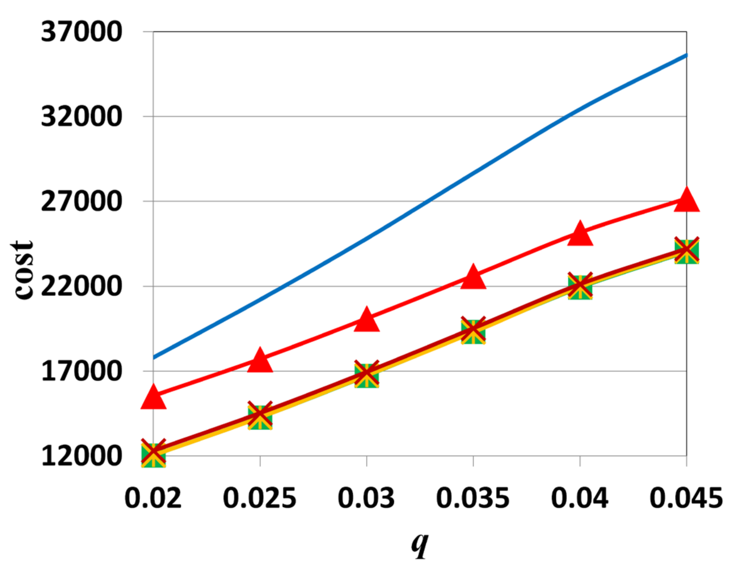


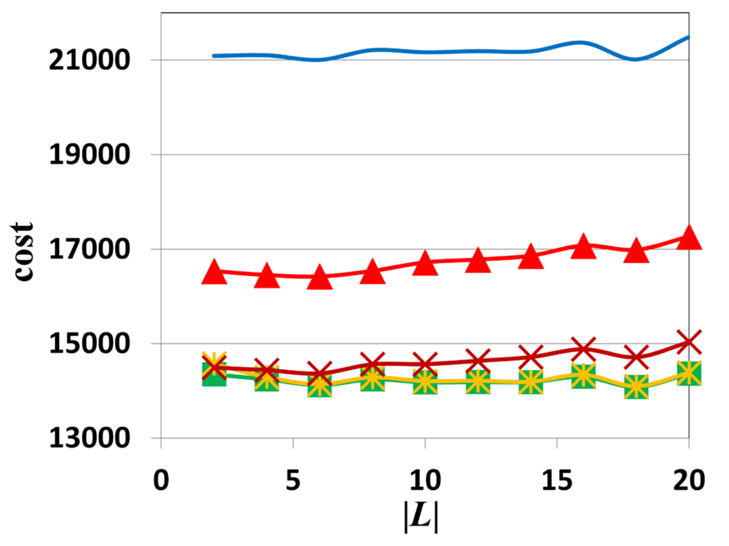
Results:

Evaluation against a baseline (B) that clusters the graph ignoring colors.









Accuracy on **synthetic datasets** in terms of similarity with respect to ground truth (***F***) and solution **cost**, by varying the level of noise (***q***), and the number of labels ($|L|$)

➢ AM **heuristic algorithm** that allows to choose the number of **output clusters**.

It finds a local optimum of the objective function based on the **alternating minimization** paradigm.

dataset	cost B	CB	LCB	AM
String	163 305	160 060	155 881	156 976
Youtube	23 550 213	18 956 000	22 644 858	19 670 899
DBLP	2 260 065	1 633 149	1 678 714	2 018 952

Cost of algorithms on **real datasets** in different domains: biological (String), social network (Youtube) and bibliographic (DBLP)

Paper ID: 531