Signed graphs: clustering and link prediction

Géraud Le Falher — MAGNET

January 22, 2015

Outline

Applications

A major source of signed graphs are graphs of social interactions, in which we want to:

- find antagonistic groups in signed graphs or in users/items bipartite graphs (Youtube, Amazon, etc) (Ailon, Avigdor-Elgrabli, *et al.* 2012)
- predict sign of unknown links (Leskovec *et al.* 2010), for instance to improve recommendation relevance

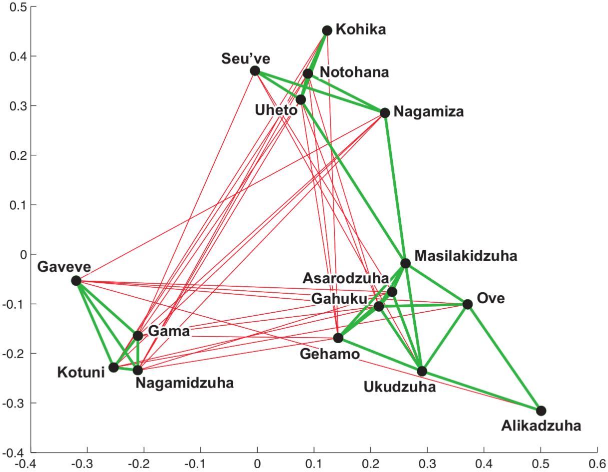


Figure: Friendly and antagonistic relations between 16 New Guinean tribes, belonging to three higher order groups found by ethnological observations (Luca *et al.* 2010)

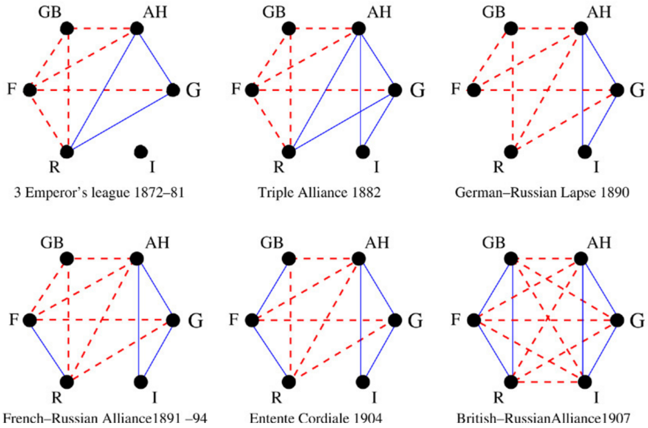


Figure: Military alliances between European states before WW1 (Antal *et al.* 2006)

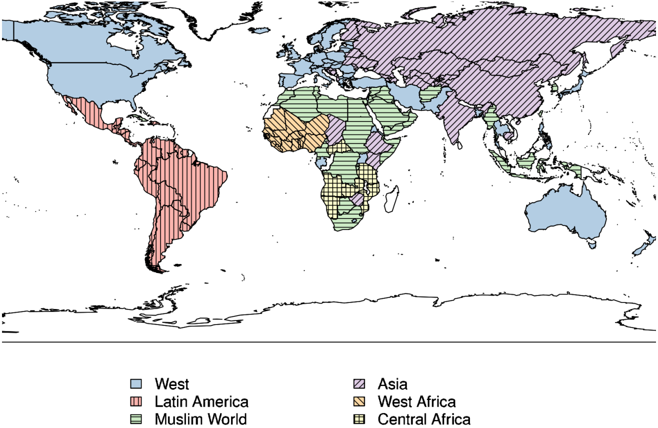
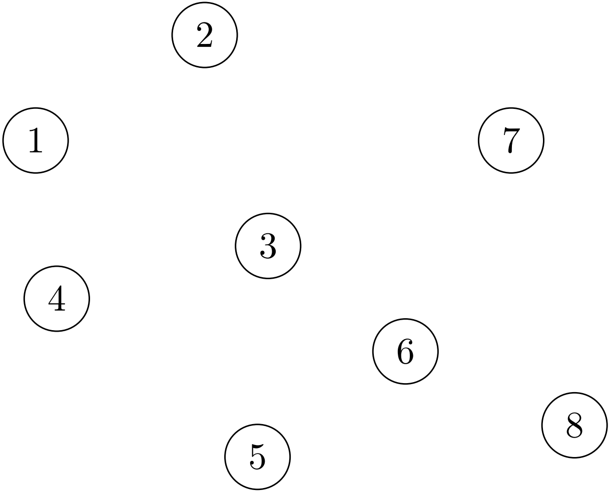


Figure: Correlates of war between 1993 and 2001, somehow reflect Huntington blocks (Traag *et al.* 2009)

The CORRELATION CLUSTERING problem (Bansal *et al.* 2002)

input

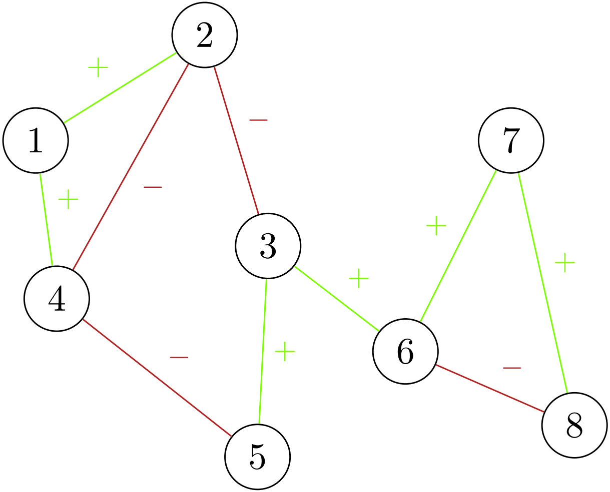
- n objects



The CORRELATION CLUSTERING problem (Bansal *et al.* 2002)

input

- n objects
- binary relation between (some of) them



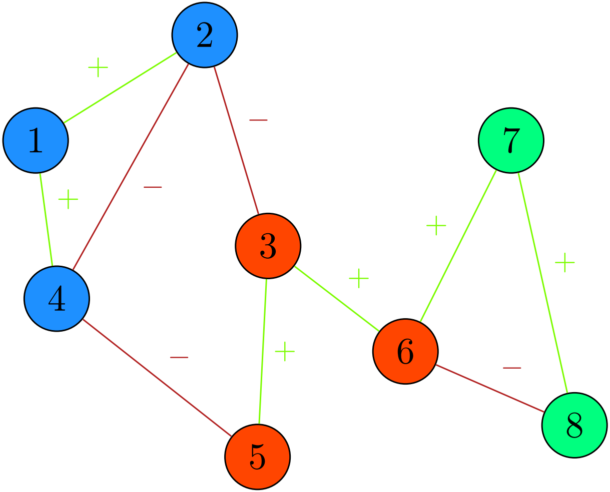
The CORRELATION CLUSTERING problem (Bansal *et al.* 2002)

input

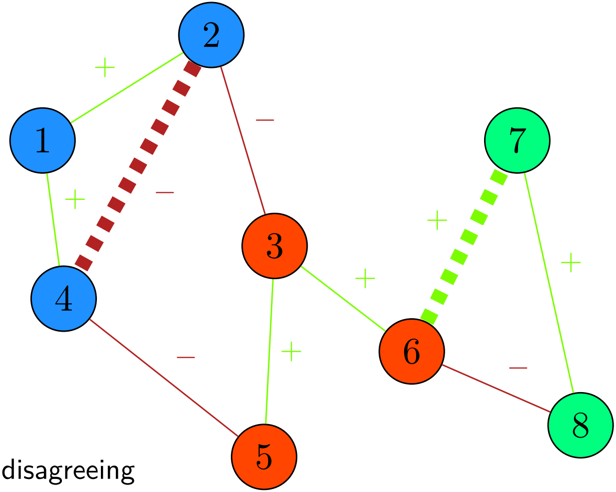
- *n* objects
- binary relation between (some of) them

output

clustering



The CORRELATION CLUSTERING problem (Bansal *et al.* 2002)



input

- n objects
- binary relation between (some of) them

output

clustering

measure of quality

- Some edges are disagreeing
- we want to minimize their number

State of the art

Two main approaches, depending of the input

Complete graph

- NP-complete by reduction from the multicut problem (Demaine *et al.* 2006)
- There is a quadratic combinatorial randomized approximation whose expected cost is at most 3 times the optimal one (Ailon, Charikar, *et al.* 2008)

State of the art

Two main approaches, depending of the input

Complete graph

- NP-complete by reduction from the multicut problem (Demaine *et al.* 2006)
- There is a quadratic combinatorial randomized approximation whose expected cost is at most 3 times the optimal one (Ailon, Charikar, *et al.* 2008)

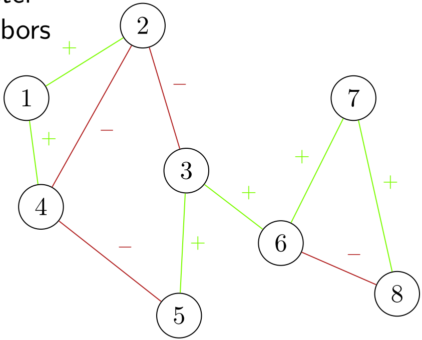
General graph

- There is a polynomial approximation (of ratio $O(\log n)$) that solves a large linear program (Demaine *et al.* 2006).
- But less information so for any constant c, getting a $O(c)$ approximation is NP-Hard.

State of the art

Complete graph

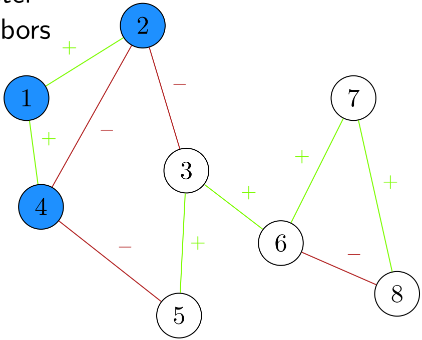
function CC-PIVOT($G = (V, E)$)
  while not all nodes are clustered **do**
    pivot $\leftarrow$ pick a node in V at random
    put *pivot* in its own cluster
    add all its positive neighbors
    remove them from G



State of the art

Complete graph

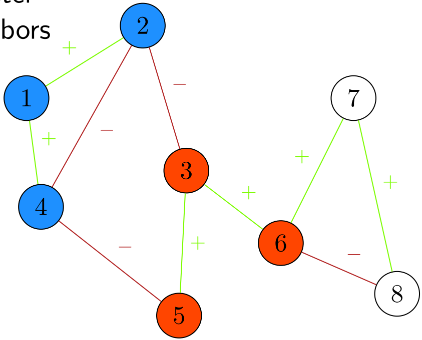
function CC-PIVOT($G = (V, E)$)
  while not all nodes are clustered **do**
    pivot $\leftarrow$ pick a node in V at random
    put *pivot* in its own cluster
    add all its positive neighbors
    remove them from G



State of the art

Complete graph

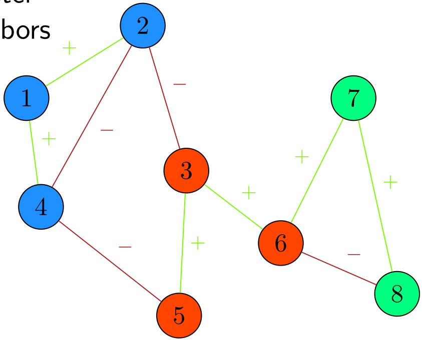
function $\textsc{CC-Pivot}(G = (V, E))$
  while not all nodes are clustered **do**
    pivot $\leftarrow$ pick a node in V at random
    put *pivot* in its own cluster
    add all its positive neighbors
    remove them from G



State of the art

Complete graph

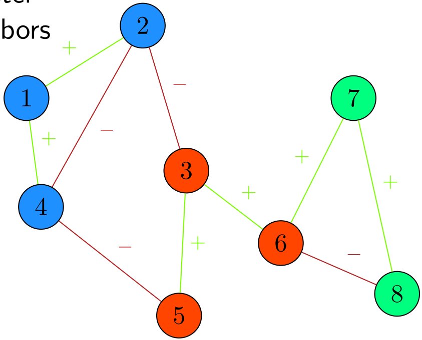
function $\textsc{CC-Pivot}(G = (V, E))$
  while not all nodes are clustered **do**
    pivot $\leftarrow$ pick a node in V at random
    put *pivot* in its own cluster
    add all its positive neighbors
    remove them from G



State of the art

Complete graph

function $\text{CC-PIVOT}(G = (V, E))$
  while not all nodes are clustered **do**
    pivot $\leftarrow$ pick a node in V at random
    put *pivot* in its own cluster
    add all its positive neighbors
    remove them from G

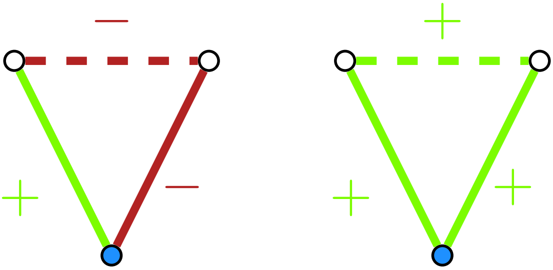


Solving the linear program brings a 2.5 approximation

Our method for general graph

idea

- complete the graph in a combinatorial fashion
- run CC-PIVOT
- keep the clustering induced on the original graph



Ongoing work

goals

- reasonable polynomial complexity
- $O(\log n)$ approximation in the worst case
- better for "realistic average-case" (Makarychev *et al.* 2014)

means

- A crucial point is how to choose the pivot for completing
- Experimental evaluation of several strategies
- Analysis on simple cases

References I

N. Ailon, N. Avigdor-Elgrabli, *et al.*, “Improved Approximation Algorithms for Bipartite Correlation Clustering”, *SIAM Journal on Computing*, vol. 41, no. 5, 2012.

N. Ailon, M. Charikar, *et al.*, “Aggregating inconsistent information”, *Journal of the ACM*, vol. 55, no. 5, 2008.

T. Antal *et al.*, “Social balance on networks: The dynamics of friendship and enmity”, *Physica D: Nonlinear Phenomena*, vol. 224, no. 1-2, 2006.

N. Bansal *et al.*, “Correlation clustering”, *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, 2002.

References II

- E. D. Demaine *et al.*, "Correlation clustering in general weighted graphs", *Theoretical Computer Science*, vol. 361, no. 2-3, 2006.
- J. Leskovec *et al.*, "Predicting positive and negative links in online social networks", in *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010.
- E. W. D. Luca *et al.*, "Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization", in *Proceedings of the 2010 SIAM International Conference on Data Mining*. 2010, ch. 48.
- K. Makarychev *et al.*, "Algorithms for Semi-random Correlation Clustering", , 2014.

References III

V. A. Traag *et al.*, "Community detection in networks with positive and negative links", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, 2009.

Thank you for your attention

Questions?

Linear Program

$$\min \sum_{(i,j)\in E^+} (1 - x_{ij}) w_{ij} + \sum_{(i,j)\in E^-} x_{ij} w_{ij}$$

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} \in [0,\ 1]$$