Thesis notes

13th July

The Echo Chamber Problem

Goal: given an interaction graph G, find $U \subseteq V$ maximing

$$\xi(U) = \sum_{C \in \hat{\mathcal{C}}} \sum_{T[U] \in S_C(U)} (|T^+[U]| - |T^-[U]|) \tag{1}$$

where $|T^-[U]|$ and $|T^+[U]|$ denotes the number of negative and positive edges induced in the subgraph, respectively.

The set of users maximing the expression is denoted as $\hat{U}$ and the corresponding score is $\xi(G)$

Purity scores

New score for evaluating our method

$$\text{Purity}(U) = \frac{\#\text{ nodes with majority label}}{|U|} \tag{2}$$

Evaluation algorithm

Evaluation algorithm for a graph $G = (V, E)$ with $\mathcal{I}$ communities.

Algorithm 1: Clustering process

foreach $i \in \mathcal{I}$ **do**

$U \leftarrow$ solve ECP on G ;

// Remove edges contributing to $\xi(U)$;

$E \leftarrow E \setminus \{e_{ij} \in E_k, T_k \in \mathcal{S}_C(U), C \in \hat{\mathcal{C}}\}$;

$C_U \leftarrow$ components of $G[U]$ considering only positive edges;

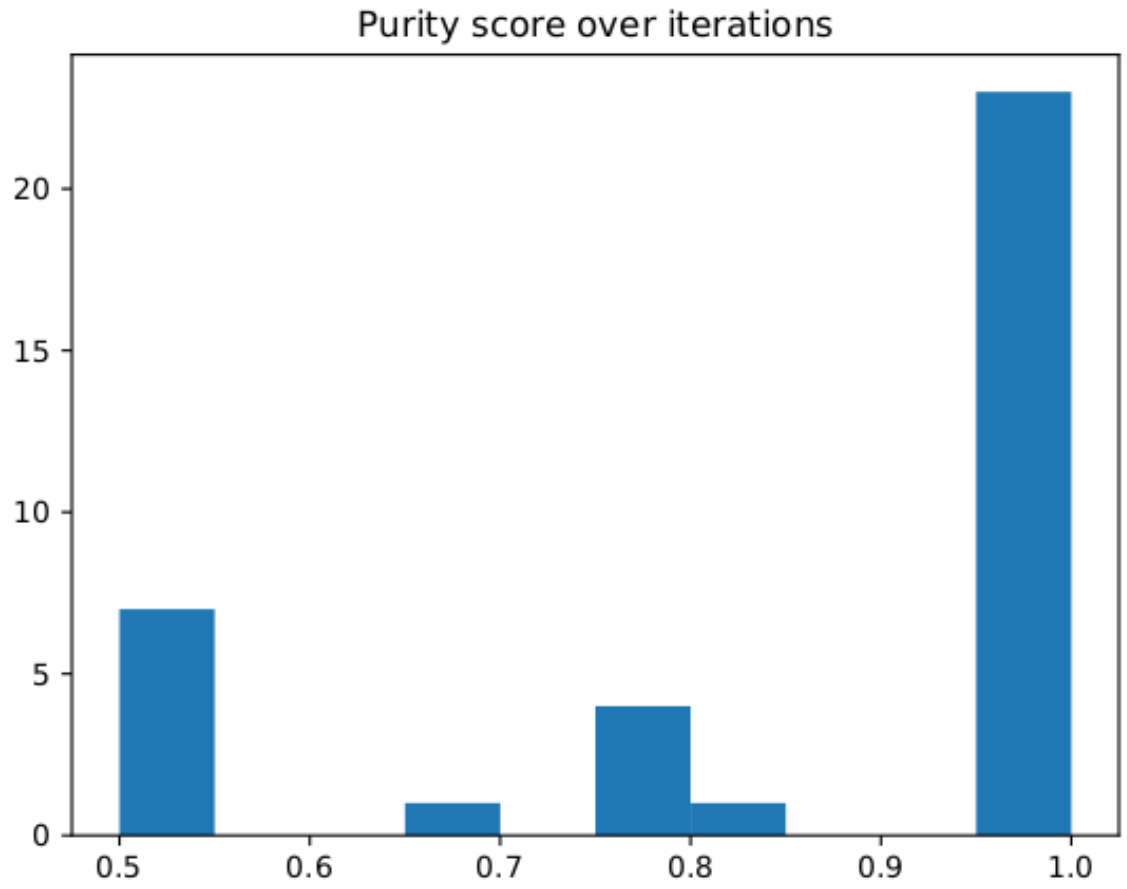
$l_j \leftarrow$ majority label of users C_j in $\mathcal{L}$, $\forall C_j \in C_U$;

$P_j = \text{Purity}(C_j)\ \forall C_j \in C_U$

end

Scores of @nytimes

Dataset has 80% with one label and the remaining 20% having another label



Results analysis

- Most of the components with scores 0.5 or 1 have two vertices
- Fraction of components with purity scores of 1 are similar to the fraction of pure components if randomly sampling two vertices from the graph

Improved Twitter labeling (1)

Based on "Birds of a Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data" by Barberá
Parameters of the model

- ϕ_i and θ_j idealogical dimension of user i and politician j. $\phi_i, \theta_j \in R$
- α_i political interest of user i
- β_j popularity of politician j

Improved Twitter labeling (2)

1. Obtain set of users O as in the paper
   - Start from politicians P and mine followers
   - Exclude inactive users and bots
2. Consider subset of O following at least 10 politicians
3. Use them to estimate parameters indexed by j

Improved Twitter labeling (2)

1. Obtain set of users O as in the paper
   - Start from politicians P and mine followers
   - Exclude inactive users and bots
2. Consider subset of O following at least 10 politicians
3. Use them to estimate parameters indexed by j
4. Use these parameters to fit parameters of users in U
5. For each user $u \in U$, $label_u =$ democrat if

$$label_u = \begin{cases} democrat, & \text{if } \phi_u < 0 \\ republican, & \text{otherwise.} \end{cases} \tag{3}$$

(during parameters initialization, democrats $\theta_j = -1$ and republicans $\theta_j = 1$).

Improved Twitter labeling (3)

Possible variant: take into account also political involvement of user α_i and exclude users with low involvement.