Thesis notes

27th April

The Echo Chamber Problem - notation

- $G = (V, E^+, E^-)$ interaction graph
- $\mathcal{C}$ set of contents
- $C \in \mathcal{C}$ content, $\mathcal{T}_C$ set of threads associated with C. A thread $T \in \mathcal{T}_C$ is a subgraph of G
- $U \subseteq V$ subset of users, $T[U]$ subgraph of T induced by U. $|T(U)|$ is the number of edges of this subgraph

The Echo Chamber Problem - notation

- $\eta(C)$ fraction of negative edges associated with C (analogous definition for a thread T). Content (or thread) controversial if $\eta \in [\alpha, 1]$
- $\hat{\mathcal{C}} \subseteq \mathcal{C}$ set of *controversial* contents
- $\mathcal{S}_C(U)$ set of *non controversial* threads induced by U, for *controversial* contents, i.e.

$$\mathcal{S}_C(U) = \{T[U] \; s.t. \; T[U] \; non \; controversial, T \in \mathcal{T}_C, C \in \hat{\mathcal{C}}, U \subseteq V\} \tag{1}$$

The Echo Chamber Problem

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

$$\xi(U) = \sum_{C \in \hat{\mathcal{C}}} \sum_{T[U] \in S_C(U)} |T[U]| \tag{2}$$

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

The Densest Echo Chamber Problem

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

$$\psi(U) = \sum_{C \in \hat{\mathcal{C}}} \sum_{T[U] \in S_C(U)} \frac{|T[U]|}{|U|} \tag{3}$$

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

A solvable Densest Echo Chamber problem (1)

Let $G = (V, E)$ be the interaction graph, $\delta(i, j)$ and $\delta^-(i, j)$ the sum of the edges and negative edges, respectively, between vertices v_i and v_j associated to controversial contents.

The graph $G_d = (V_d, E_d)$ is constructed as follows from G:

- for any vertex $v_i \in V$ add a corresponding vertex in V_d
- for any pair of vertices in G
  - let $\eta(i, j) := \frac{\delta^-(i,j)}{\delta(i,j)}$. If $\eta(i, j) \leq \alpha$ add a positive edge between v_i and v_j in G_d

Densest non controversial subgraph:
Let $E_d[U]$ the set of edges induced on G_d by $U \subseteq V$. **Goal**: find U maximizing

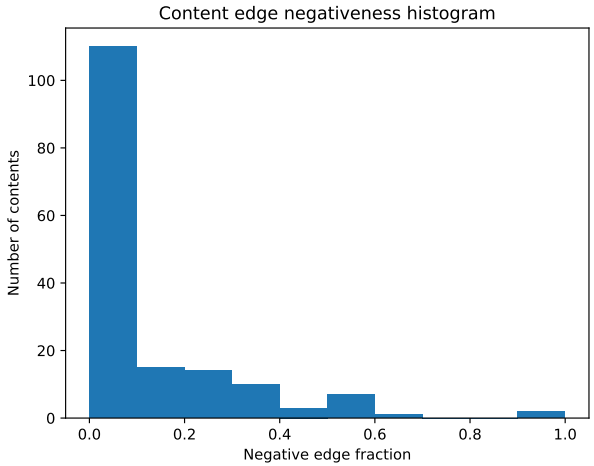
$$\xi(U) = \frac{|E_d[U]|}{|U|} \tag{4}$$

A solvable Densest Echo Chamber problem (2)

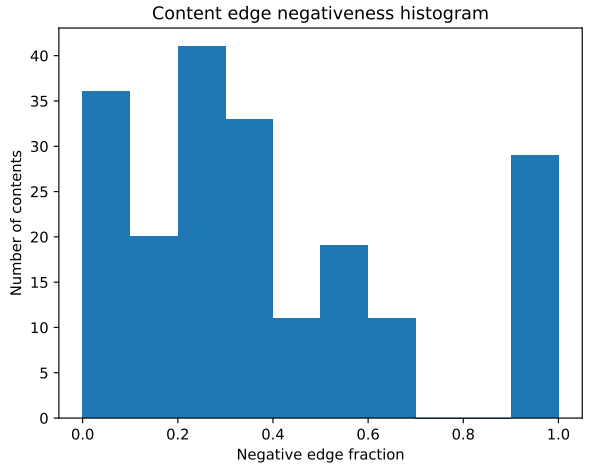
Alternative taking into account threads:

- T-Densest non controversial subgraph: aggregate edges separately for each $T \in \mathcal{T}_C, C \in \hat{\mathcal{C}}$, i.e. let $\delta_T(i,j)$ be $\delta(i,j)$ for the subgraph T:
  - let $\eta(T,i,j) := \frac{\delta_T^-(i,j)}{\delta_T(i,j)}$. If $\eta(T,i,j) \leq \alpha$ add a positive edge

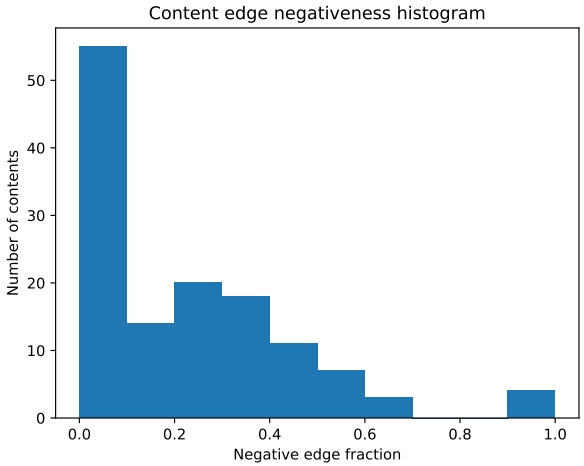
The datasets - negative edge fractions for contents



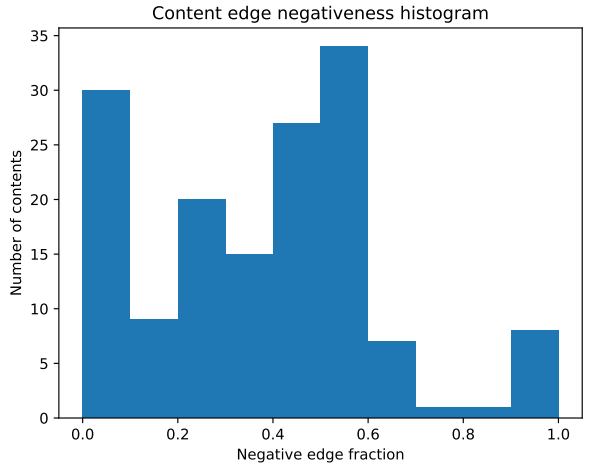
(a) r/cats



(b) r/covid19

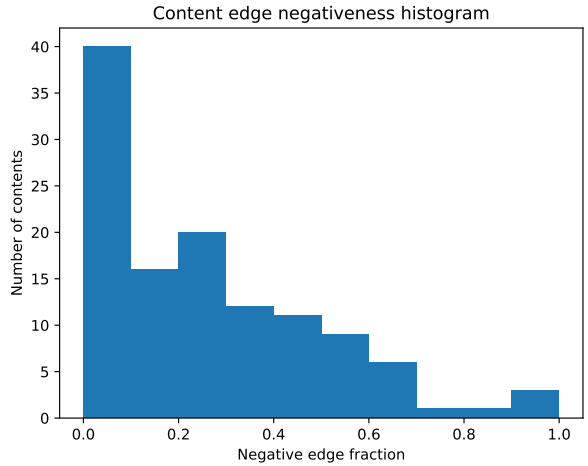


(c) r/programming

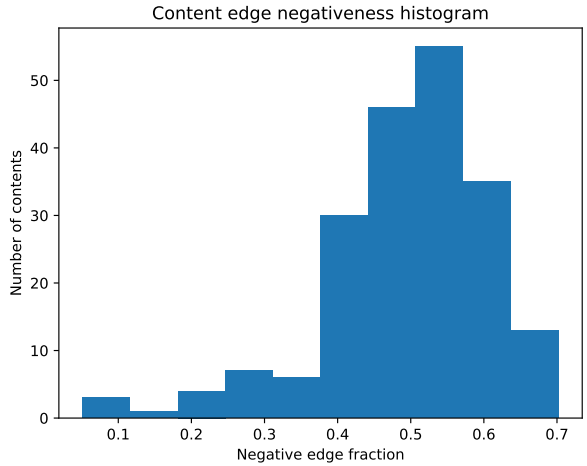


(d) r/climate

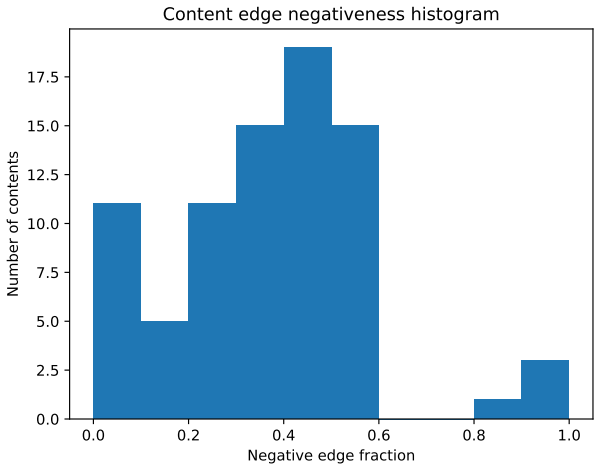
The datasets - negative edge fractions for contents



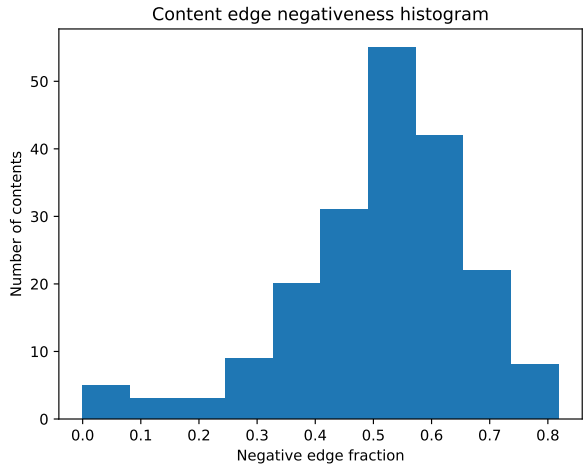
(a) r/football



(b) r/asktrumpsupporters



(c) r/economics



(d) r/politics

Results

Table: Echo chamber scores. For the $\xi_{round}(G)$ the results tuple corresponds to (score, $|U|$, number of contributing threads, time in seconds). In the other scores the number of contributing thread is omitted. $\alpha = 0.4$

Source, $\lvert V\rvert$, $\lvert E\rvert$, $\lvert\hat{\mathcal{C}}\rvert$	$\xi_{round}(G)$	$\psi_{Dens}(G)$	$\psi_{T-Dest}(G)$
cats, 2162, 2866, 11	(16, 27, 4, 1)	(0.75, 7, 12)	(0.75, 16, 12)
covid19, 2595, 5806, 65	(295, 216, 28, 4)	(1.14, 7, 17)	(1.57, 16, 17)
programming, 3728, 6453, 25	(836, 572, 36, 10)	(0.92, 15, 34)	(0.92, 15, 34)
climate, 6457, 13150, 74	(2719, 1488, 140, 102)	(1.30, 20, 102)	(4.5, 2, 105)
football, 6407, 10498, 28	(3147, 1774, 43, 125)	(1, 9, 102)	(1, 9, 104)
asktrumpsupporters, 8316, 62449, 172	(8716, 3023, 100, 29741)	(5.3, 185, 173)	(12, 2, 179)
economics, 9871, 19281, 36	(9545, 4915, 85, 1042)	(1.44, 29, 241)	(1.47, 37, 247)
politics, 28695, 59334, 159	(36696, 17420, 243, 16438)	(2.49, 55, 1967)	(5.33, 3, 1940)

A solvable Densest Echo Chamber problem (3)

Other alternatives:

- Compute DCS-** on G, where each snapshot corresponds to a content
  - DCS-MM and DCS-MA trivially 0 due to sparseness
  - DCS-AA equivalent to previous measure
  - DCS-AM very close to 0 due to sparseness
- Solve O^2Bff problem on the preprocessed graph sequence, i.e.

  Given a graph history $\mathcal{G} = \{G_1, G_2, ..., G_\tau\}$, an aggregate density function f , and an integer k, find a subset of nodes $S \subseteq V$, and a subset $\mathcal{C}_k$ of $\mathcal{G}$ of size k, such that $f(S, C_k)$ is maximized

  O^2Bff-AM can be computed and more meaningful for a good choice of k.

A model for the Echo Chamber Problem (1)

Model parameters:

- b_i, the group of each user i
- ω_{rs}^{+} and ω_{rs}^{-}, the probabilities of positive and negative edges, respectively, between users in group r and s ($\omega_{rs}^{+} + \omega_{rs}^{-} \leq 1$).
- θ, controlling the reduction of probability of interacting between *inactive* communities

For each content:

1. Sample n' among the n communities. These are the *active* communities in the content discussion
2. For each node pairing i, j consider their corresponding groups r and s.
   - If both communities are *active* draw from the categorical distribution $(\omega_{rs}^{+}, \omega_{rs}^{-}, 1 - \omega_{rs}^{+} - \omega_{rs}^{-})$ to add an edge (or not).
   - Otherwise draw from the categorical $(\theta\omega_{rs}^{+}, \theta\omega_{rs}^{-}, 1 - \theta(\omega_{rs}^{+} + \omega_{rs}^{-}))$, $\theta \leq 1$.

A model for the Echo Chamber Problem (2)

Again each node has a group assignment and there are probabilities of positive and negative edges ω_{rs}^{+} and ω_{rs}^{+}, respectively.

1. Generate the *follow* graph G by using a SBM with parameters $\{\phi_{rs}\}$.
2. Each node can be active with probability β_a
3. Any active node activates his inactive neighbours in G with probability β_n
4. active nodes interact according to the categorical $(\omega_{rs}^{+}, \omega_{rs}^{-}, 1 - \omega_{rs}^{+} - \omega_{rs}^{-})$ otherwise (at least one of the 2 nodes is inactive) with categorical $(\theta\omega_{rs}^{+}, \theta\omega_{rs}^{-}, 1 - \theta(\omega_{rs}^{+} + \omega_{rs}^{-}))$, $\theta \leq 1$

Computing the score on the synthetic data (1)

The following results are based on the second model. Results are also evaluated against known nodes communities by repeatedly solving the echo chamber problem on the graph (from which edges contributing to the score are iteratively removed).

Clustering is evaluated through Rand score, i.e. the fraction of pair of nodes which correspond in the predicted and true clustering. These are:

- pairs of nodes which are in the same cluster in the predicted and true clustering;
- pairs of nodes which are in different clusters in the predicted and true clustering.

Also an *adjusted* version of the measure is provided, comparing the score against a random assignment of the labels.

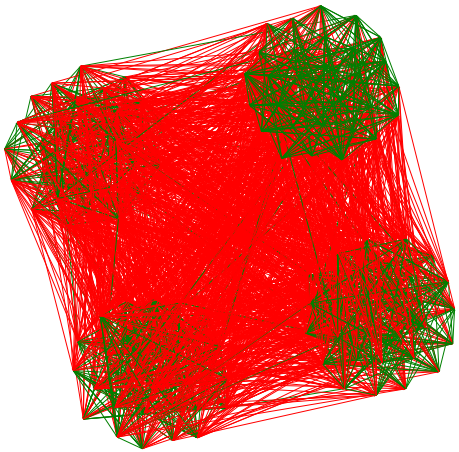
Computing the score on the synthetic data (2)

The following graphs contains 4 communities among which vertices are equally distributed and 10 threads, with $\alpha = 0.2$. The results have been computed with the non-exact algorithm.

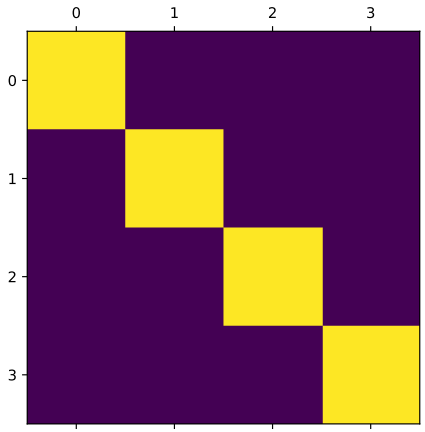
$\beta_{\boldsymbol{a}} = 1/8$, $\beta_{\boldsymbol{n}} = 1/3$, $\theta = 1/10$, $\phi = 0.7$ (usually) for nodes inside the same community and $\phi = 0.2$ for nodes in different community. These parameters control graph density.

Computing the score on the synthetic data (3)

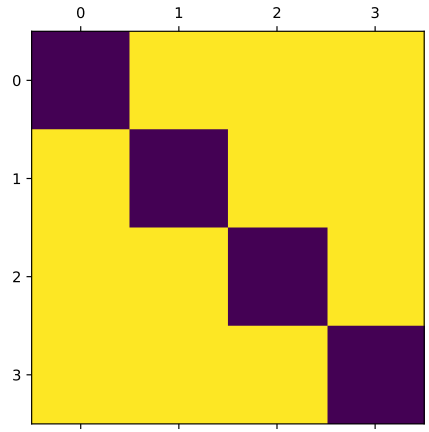
First graph: clean division of communities with mainly positive edges inside the groups and negative edges between groups.



(a) Graph



(b) ω_{rs}^{+}

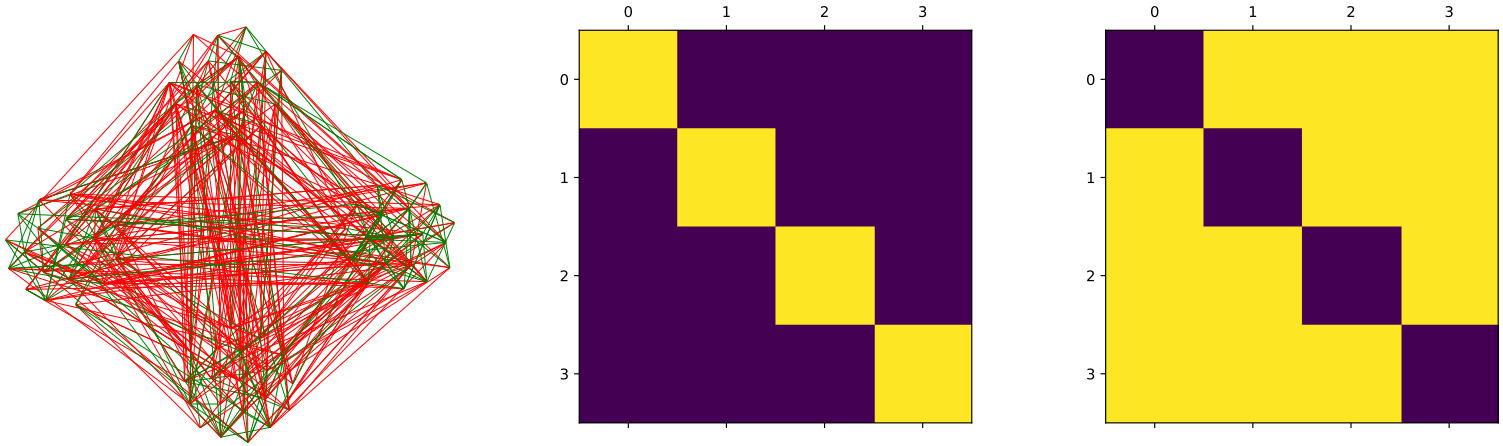


(c) ω_{rs}^{-}

$|V| = 120,\ |E| \approx 2400,\ \eta(G) \approx 0.6,\ \bar{\xi}(G) \approx 91.$
Rand $= 0.7$, adjusted Rand $= 0.25$

Computing the score on the synthetic data (4)

Second graph: much smaller distinction in the interaction between communities.

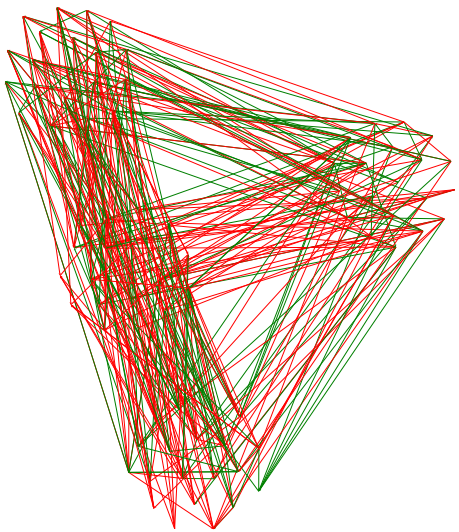


(a) Graph (b) ω_{rs}^{+} (c) ω_{rs}^{-}

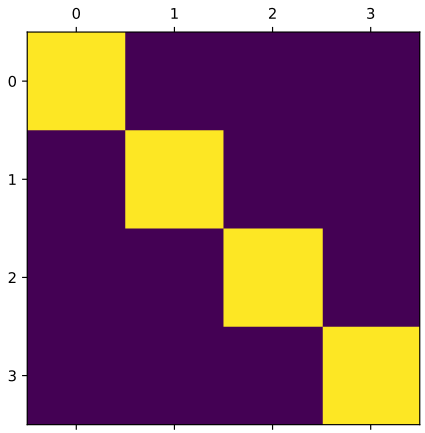
$|V| = 80,\ |E| \approx 400,\ \eta(G) \approx 0.7,\ \bar{\xi}(G) \approx 16.$
Rand $= 0.6$, adjusted Rand $= 0.026$

Computing the score on the synthetic data (5)

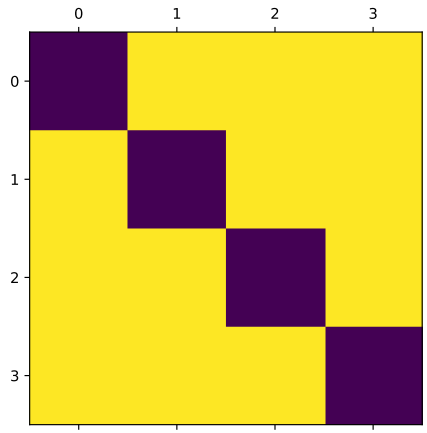
Third graph: much smaller distinction in the interaction between communities.



(a) Graph



(b) ω_{rs}^{+}



(c) ω_{rs}^{-}

$|V| = 80,\ |E| \approx 400,\ \eta(G) \approx 0.58,\ \bar{\xi}(G) \approx 25.$
Rand $= 0.6$, adjusted Rand $= 0.007$

Observations on the result

- Increasing the number of vertices in each community helps clustering correcly the nodes
- the lower the distiction of interactions between communities, the more it is difficult to find a correct clustering
- in order to find a good clustering α may be chosen to be equal to the η inside each of the communities, so that a single whole community does not produce many controversial threads in the induced subgraphs