Thesis notes

4th May

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 $T \in T_C, C \in \hat{\mathcal{C}}$, 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

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

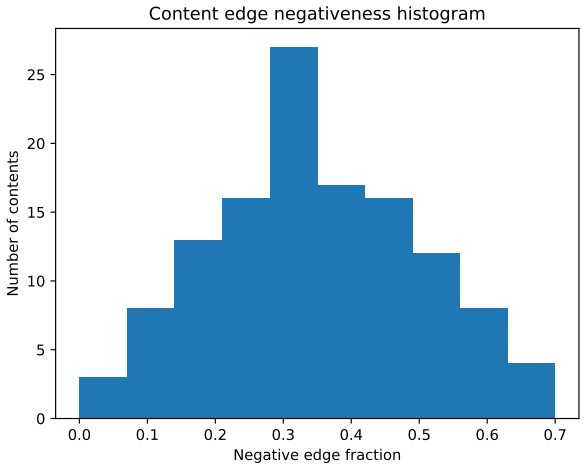
O^2Bff-AM can be computed.

Results

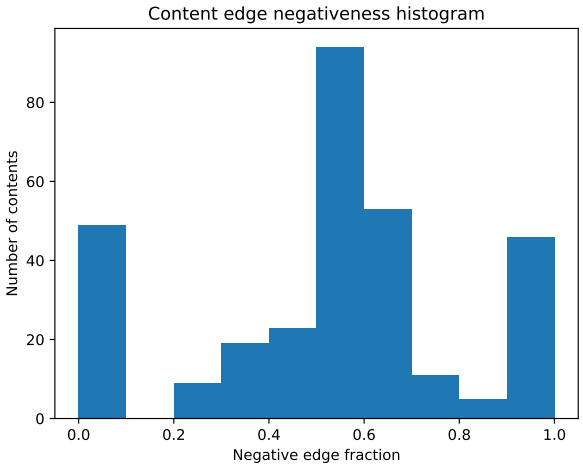
The algorithm (even if it is just an approximation) requires a lot of times even on smaller graphs.

For a graph with 100 snapshots and $k = 10$, $|V| \approx 2000$ and $|E| \approx 2000$ it needs > 6 hours.

Results



(a) @nytimes



(b) @foxnews

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, $\|V\|$, $\|E\|$, $\|\hat{\mathcal{C}}\|$	$\xi_{round}(G)$	$\psi_{Dens}(G)$	$\psi_{T-Dest}(G)$
nytimes, 20051, 24468, 44	(5961, 4814, 243, 400)	(1.2, 16, 1000)	(1.2, 16, 1000)
foxnews, 45509, 82494, 232	(50240, 26352, 1090, 24000)	(1.96, 28, 5000)	(1.97, 35, 5000)

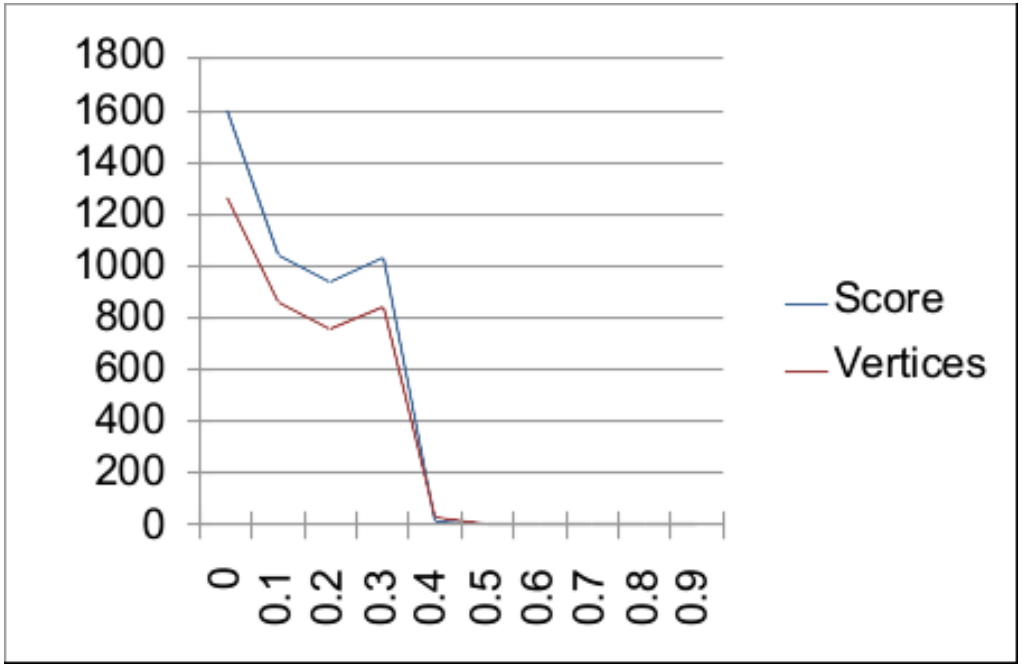
Table: Echo chamber scores for alpha chosen as the median of the η of the contents. Tuple meaning is the same as above

Source, $\|V\|$, $\|E\|$, $\|\hat{\mathcal{C}}\|$	$\xi_{round}(G)$	$\psi_{Dens}(G)$	$\psi_{T-Dest}(G)$
nytimes, 20051, 24468, 62	(7337, 5775, 288, 243, 400)	(1.2, 16, 1000)	(1.2, 16, 1000)
foxnews, 45509, 82494, 154	(53643, 30816, 810, 24000)	(1.96, 28, 5000)	(1.97, 35, 5000)

Median alpha for @nytimes: 0.358
Median alpha for @foxnews: 0.553

Choosing alpha

It could be interesting to indagate the relationship between the Echo Chamber Score and alpha for different datasets, also relating it to the distribution of η for the contents and threads.



A model for the Echo Chamber Problem

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 $(\theta\omega_{rs}^{+}, \theta\omega_{rs}^{-}, 1 - \theta(\omega_{rs}^{+} + \omega_{rs}^{-}))$, $\theta \leq 1$

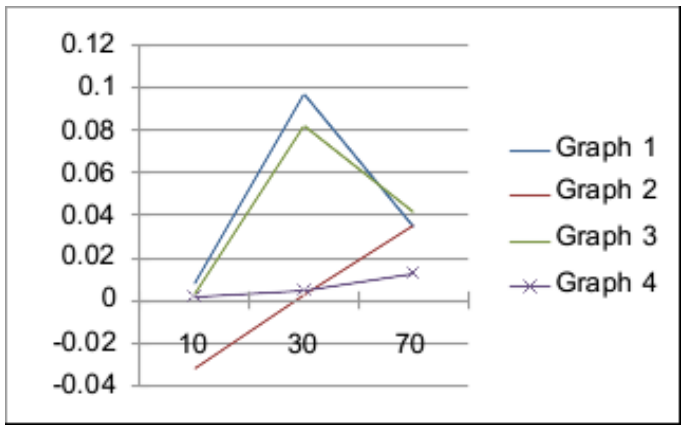
Computing the score on the synthetic data (1)

Results are 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).
4 different graphs are reported, graph n having less "cohesive" communities than graph $n - 1$.

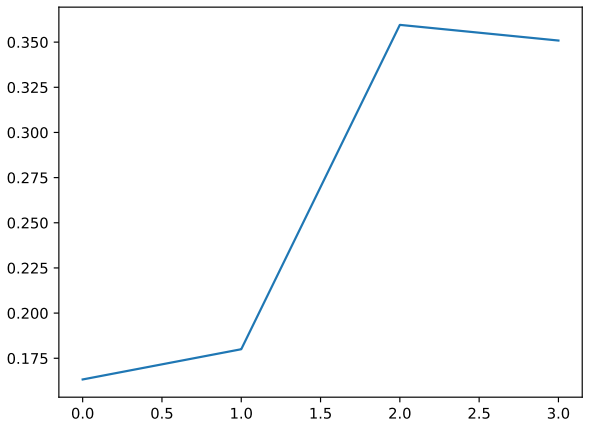
In each graph there are 4 communities and 10 threads; Adjusted rand index is plotted to measure final clustering, Jaccard for measuring new cluster similarity along the iterations

Computing the score on the synthetic data (2)

Test 1: nodes uniformly distributed among communities, $\alpha = 0.2$.



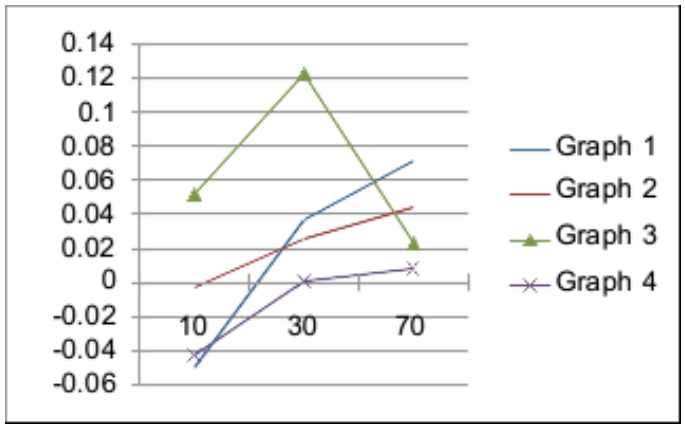
(a) Adj RAND



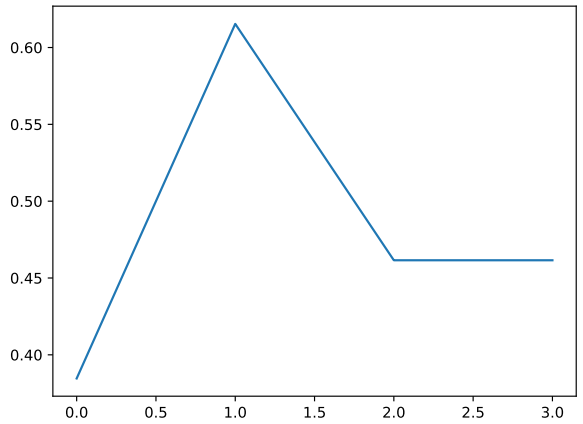
(b) Jaccard along iterations, for (Graph 1, 30)

Computing the score on the synthetic data (2)

Test 2: 2 communities have n nodes, the other $n/4$, $\alpha = 0.2$.



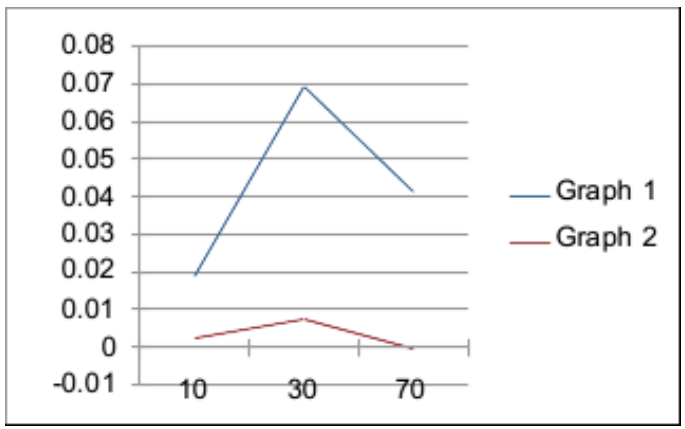
(a) Adj RAND



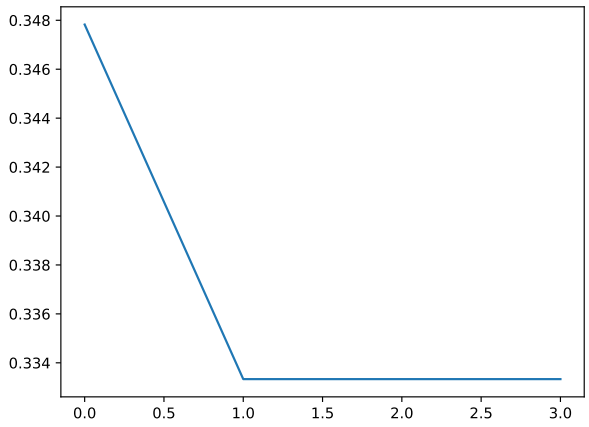
(b) Jaccard along iterations, for (Graph 3, 10)

Computing the score on the synthetic data (2)

Test 3: 2 communities have n nodes, the other $n/4$, α chosen as the ratio of probability of negative edge over probability of edge inside the communities.



(a) Adj RAND



(b) Jaccard along iterations, for (Graph 1, 10)

Observations on the result

Possible reasons for the poor results

- Solution of the Echo Chamber problem may find smaller groups of users inside the communities
- choice of α
- results are also noisy due to the approximate algorithm