Dynamics of opinion polarization

Elisabetta Biondi, Chiara Boldrini, Andrea Passarella, Marco Conti

S. Iochanas T. Lagkalis

Technical University of Crete,
Large and Social Networks: Modeling and Analysis (TEL422)

Table of Contents

The Problem: How do people form opinions?

- The rise of online social media has played a big role on the formation of peoples' opinion.
- The provide an algorithm personalization which reinforces cognitive biases, reducing discomfort experienced when exposed to opposite opinions and thus, creates **echo chambers**.
- Whether this is the reason of polarized opinions is still debatable in literature. Some argue that some inherent characteristics of social nets and human interaction are predominant than algorithmic filtering.

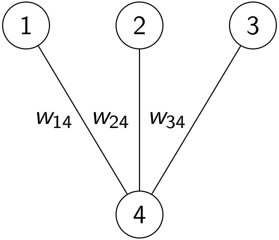
The importance of the problem

- Opinion polarization has a huge social relevance.
- It's a scientific challenge, since only a few models can describe opinion dynamics while being both realistic and mathematically tractable at the same time.
- This paper's focus is the investigation of opinion dynamics through the **Friedkin-Johnsen (FJ) model**, which incorporates individual stubbornness (resistance to change of their opinion).
- Explores whether this model can actually explain polarization and under what conditions.

Table of Contents

Related Work: DeGroot's model

Opinion of a node at each (discrete) time step is the average opinion of it's neighbor nodes, weighted by the strength of their social influence.



$$z_4(t+1) = \sum_j w_{4j} z_j(t)$$

where $z_i(t)$ is the opinion of node i at timestep t, w_{ij} is the influence of node i to node j.
When it converges all nodes have the same opinion (consensus) and no polarization $\implies$ not realistic.

Related Work on Friedkin - Johnsen model

Generalization of DeGroot's model. Introducing λ_i - **susceptibility**, i.e. willingness of a node to accept new opinions.

$$z_4(t+1) = (1-\lambda_4)z_4(0) + \lambda_4 \sum_j w_{4j} z_j(t)$$

Pros:

- Better representation of opinion dynamics

Cons:

- It's not clear if captures polarization of not!

This paper derives the conditions under which FJ model yields polarization.

Table of Contents

Problem formulation

- The paper focuses on undirected graphs representing social networks. Two types of graphs with the same vertices $\mathcal{V}$ and edges $\mathcal{E}$:
  1. Social graph, $\mathcal{S}$ where each edge weight is the number of social interactions.
  2. Influence graph, $\mathcal{I}$ where each edge weight w_{ij} is the influence of node i to node j.

  $\mathcal{I}$ can be derived from $\mathcal{S}$, since stronger social relationships will influence more than weak ones.
- Weights are normalized: $w_{ij} = \frac{\hat{w}_{ij}}{\sum_{j=1}^{n} \hat{w}_{ij}}$
- $W = (w_{ij})$ is the influence matrix. It's row stochastic (i.e. rows sum to 1)
- Time is discrete.
- Opinion of node i at time k: $z_i(k)$. In general $z_i(k) \in \mathbb{R}$ but in this paper is assumed $z_i(k) \in [-1, 1]$.

Problem formulation (cntd.)

- $N(i)$ is the neighborhood of node i.
- s_i is the initial opinion (prejudice) of node i ($s_i = z_i(0)$).
- $\lambda_i \in [-1, 1]$ susceptibility if node i, i.e. willingness of a node to accept new opinions.
- Convergence: $\forall i z_i(k+1) \to z_i$ as $k \to \infty$
- Consensus: $\forall i z_i(k+1) \to z$ as $k \to \infty$
- A **choice shift** occurs when the mean attitude of the group at the end is different from the mean attitude at the beginning:

$$\sum_i z_i \neq \sum s_i$$

Table of Contents

FJ family of models

Table: The FJ family of models

Model	Update Equation
Generalized FJ (gFJ)	$z_i(k+1) = (1-\lambda_i)s_i + \lambda_i \sum_{j \in \{i\} \cup N(i)} w_{ij} z_j(k)$
Variational FJ (vFJ)	$z_i(k+1) = \frac{\hat{w}_{ii} s_i + \sum_{j \in N(i)} \hat{w}_{ij} z_j(k)}{\hat{w}_{ii} + \sum_{j \in N(i)} \hat{w}_{ij}}$
Restricted FJ (rFJ)	$z_i(k+1) = \frac{s_i + \sum_{j \in N(i)} \hat{w}_{ij} z_j(k)}{1 + \sum_{j \in N(i)} \hat{w}_{ij}}$

- **vFJ** excludes the agent's own current opinion from the update.
- **rFJ** is the simplified version of FJ models. It sets the weight of the internal opinion to 1 ($w_{ii} = 1$), allowing only indirect control over susceptibility through the social weights, and is widely chosen in literature for its mathematical tractability.

Table: Matrix H for the different FJ models

Model	Matrix Formulation (H)
Generalized FJ (gFJ)	$H_g = (I - \Lambda W)^{-1}(I - \Lambda)$
Variational FJ (vFJ)	$H_v = (D + \tilde{A} - A)^{-1}\tilde{A}$
Restricted FJ (rFJ)	$H_r = (D + I - A)^{-1}$

where Λ is the diagonal susceptibility matrix.

- In each model the final opinion vector $\mathbf{z}$ can be calculated as $\mathbf{z} = H\mathbf{s}$ where H is a matrix that varies depending on the specific FJ version considered (see table 2).
- The equation of H_g holds and gFJ model is convergent and its unique stationary point $\mathbf{z}$ (i.e., the steady-state opinion vector) iff ΛW is stable (i.e., it has eigenvalues inside the open unit circle $\{z \in \mathbb{C} : |z| < 1\}$)

Polarization metrics

Polarization

For a polarization index Φ, we say that an opinion formation model $\mathcal{M}$ is Φ-polarizing (or simply polarizing for Φ) if there exists at least one initial opinion vector $\mathbf{s}$ such that the corresponding final opinion vector $\mathbf{z}$ satisfies:

$$\Phi(\mathbf{z}) > \Phi(\mathbf{s}).$$

The induced polarization is measured by the polarization shift:

$$\Delta\Phi(\mathbf{s}) = \Phi(\mathbf{z}) - \Phi(\mathbf{s}).$$

Polarization indices

Let $\mathbf{x} = (x_i) \in [-1, 1]^n$ be an opinion vector.

$$\mathrm{NDI}(\mathbf{x}) = \sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2 \qquad \mathrm{GDI}(\mathbf{x}) = \sum_{i<j}(x_i - x_j)^2$$

$$P_1(\mathbf{x}) = \sum_i (x_i - \bar{x})^2 = \|\mathbf{x} - \bar{x}\|_2^2 \qquad P_2(\mathbf{x}) = \frac{1}{|V|}\sum_i x_i^2 = \frac{1}{|V|}\|\mathbf{x}\|_2^2$$

$$P_3(\mathbf{x}) = \sum_i x_i^2 = \|\mathbf{x}\|_2^2 \qquad P_4(\mathbf{x}) = \sum_i |x_i| = \|\mathbf{x}\|_1$$

Polarization invariants

Many polarization metrics exist, but they are often treated in isolation. The lemma 1-3 in the paper reveals how they relate:

- Global Disagreement Index (GDI) and P_1 are equivalent: $GDI(x) = |\mathcal{V}|P_1(x)$
- P_3 and P_2 are equivalent: $P_3(x) = |\mathcal{V}|P_2(x)$
- Polarization invariant: $P_1(\mathbf{x}) \geq P_3(\mathbf{x}) - \frac{P_4(\mathbf{x})^2}{|\mathcal{V}|}$

Corollary 1

If there's no choice shift (mean opinion stays constant), then: Polarization increases in $P_1 \longleftrightarrow$ also increases in P_2

Table: Classes of Polarization

Type	**What is Captured**	**Indices**
Local	Opinion spread among neighboring nodes	NDI
Dispersion	Opinion spread among all nodes	GDI, P_1
Absolute (Quadratic)	Closeness to extreme opinions (squared)	P_2, P_3
Absolute (Linear)	Total deviation from neutrality	P_4

gFJ is globally polarizing but locally depolarizing

- (Theorem 2) gFJ model is **NDI-depolarizing** for any initial opinion vector. Intuition of the proof: the expressed opinion $(z_i)_i$ is the Nash Equilibrium of cost function. i.e. z_i minimizes f_i, $\forall i$, so that $f_i(z_i) \leq f_i(s_i), \forall i$.
- (Theorem 3) gFJ is **globally polarizing** iff the matrix $H_g = (I - \Lambda W)^{-1}(I - \Lambda)$ is **not doubly stochastic**. If there are any **naive nodes** ($\lambda = 1$), polarization occurs. Otherwise, polarization depends on imbalance in influence/susceptibility.
- (Theorem 4) The most polarizing vectors (i.e. initial opinion vectors **s** which lead to polarization) can be chosen with **concordant signs** (all entries ≥ 0 or ≤ 0). This means that cooperation among like-minded nodes increases polarization more effectively than conflict.

gFJ (cntd.)

- (Theorem 5) Provides analytical construction of polarizing vectors using **singular value decomposition** of H_g. Shows how to compute:
  1. Local maximum vector on an l_2-Ball: $\mathbf{s}_{\mathbf{B_2(1)}}$, and scaled version: $\mathbf{s}_{\mathbf{B_2(t)}}$
  2. Global maximum (NP-Hard): $\mathbf{s}_{\mathbf{max}}^{\mathbf{P_2,P_3}}$
  3. Convex approximations: $\mathbf{s}_{>1}$ and $\mathbf{s}_{\geq 1}$
- (Corollary 3) Efficient heuristics can approximate the optimal polarizing vector when exact solution is intractable.

Metric	polarizing vector	Brief explanation
P_2, P_3	$s_{B_2(1)}$	eigenvectors with the largest eigenvalue of $H_g^T H_g$, correspond to the local maximum for the P_2, P_3 on the L2-ball of radius 1 (Th. 5)
P_2, P_3	$s_{B_2(t)}$	The same with $s_{B_2(1)}$ but in ball of radius t. Obtained with $t \cdot s_{B_2(1)}$ (Th. 5)
P_2, P_3	$s_{max}^{P_2,P_3}$	global max for these metrics. Is the solution of a non-Convex opt. problem (NP-Hard) (Th. 5).
P_2, P_3	$s_{V>1}, s_{V\geq 1}$	Heuristic-based solution on subspaces. Approximation of the global solution (Col. 3).
P_4	$s_{max}^{P_4}$	Solution of linear opt. problem (Th. 6).
P_4	$s_{B_1(1)}$	Maximum of P_4 in 1-norm ball with radius 1.
P_1, GDI	-	if gFJ is depolarizing for P_1, P_2, P_3 then also for P_1, GDI. Otherwise, polarizing under condition (Th. 7).

Table: Summarize of the Theorems.

What about vFJ and rFJ models?

- (Theorem 8) **vFJ polarizes under the exact same conditions as gFJ** (for all metrics), if self-loop weights are set correctly $\implies$ When the social graph is undirected (i.e. matrix $\hat{W}$ is symmetric), vFJ is polarizing with P_2 , P_3 and P_4 iff $\hat{w}_{ii}$ are not identical for all i.
- (Corollary 7) The rFJ model on undirected social graphs is never polarizing, in any polarization metrics, for any initial opinion vector.

Table of Contents

Paper Results - gFJ on the Karate network dataset

$\lambda_i \propto \mathrm{P}_i$

	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}	Δ_{NDI}	Δ_{GDI}
s_{unif}	$-8.8e{-1}$	$-1.7e{-2}$	$-5.7e{-1}$	$3.1e{-1}$	-1.03	-29.93
$s_{B_2(1)}$	$-6.4e{-2}$	$2.2e{-3}$	$7.6e{-2}$	$4.2e{-1}$	$-3.1e{-1}$	-2.18
$s_{B_2(t)}$	-1.15	$4.0e{-2}$	1.36	1.77	-5.61	-38.94
$s_{V_{>1}}$	-1.18	$4.1e{-2}$	1.40	1.80	-5.67	-40.28
$s^{heu}_{V_{>1}}$	-1.18	$4.1e{-2}$	1.40	1.80	-5.67	-40.28
$s_{\max P_{2,3}}$	-1.81	$6.0e{-2}$	2.05	2.05	-6.94	-61.48
$s_{B_1(1)}$	$-2.0e{-1}$	$-5.4e{-3}$	$-1.9e{-1}$	$1.6e{-1}$	$-1.3e{-1}$	-6.63
$s_{\max P_4}$	-3.57	$3.4e{-2}$	1.16	2.65	-10.65	-121.421
			$\lambda_i \propto \mathrm{P}_i^{-1}$			
s_{unif}	-1.62	$-4.8e{-2}$	-1.63	$-1.1e{-2}$	-1.41	-55.22
$s_{B_2(1)}$	$6.8e{-1}$	$4.3e{-2}$	1.45	3.32	$-3.1e{-1}$	23.19
$s_{B_2(t)}$	$7.6e{-1}$	$4.3e{-2}$	1.63	3.52	$-3.5e{-1}$	25.97
$s_{V_{>1}}$	$3.0e{-1}$	$1.3e{-1}$	4.48	7.50	-5.08	10.07
$s^{heu}_{V_{>1}}$	$3.1e{-1}$	$1.1e{-1}$	3.61	6.74	-3.72	10.45
$s_{\max P_{2,3}}$	-2.30	$2.01e{-1}$	6.86	6.86	-3.98	-78.17
$s_{B_1(1)}$	$3.1e{-1}$	$2.2e{-2}$	$7.5e{-1}$	2.98	-1.95	10.67
$s_{\max P_4}$	-4.27	$1.3e{-1}$	4.59	10.04	-9.36	-145.22
			$\lambda_i = 0.8$			
s_{unif}	-2.78	$-7.7e{-2}$	-2.62	$1.7e{-1}$	-2.45	-94.67
$s_{B_2(1)}$	$-3.3e{-1}$	$1.5e{-1}$	$5.2e{-1}$	2.44	-1.22	-11.30
$s_{B_2(t)}$	-1.43	$6.6e{-2}$	2.23	5.06	-5.22	-48.49
$s_{\max P_{2,3}}$	-1.73	$1.3e{-1}$	4.57	4.57	-3.33	-58.73
$s_{B_1(1)}$	$-8.2e{-1}$	$-1.5e{-2}$	$-5.2e{-1}$	2.34	-5.60	-27.90
$s_{\max P_4}$	-6.86	$-1.7e{-1}$	$-1.7e{-1}$	8.10	-13.32	-233.38

Paper Results - gFJ on the Karate network dataset (cntd.)

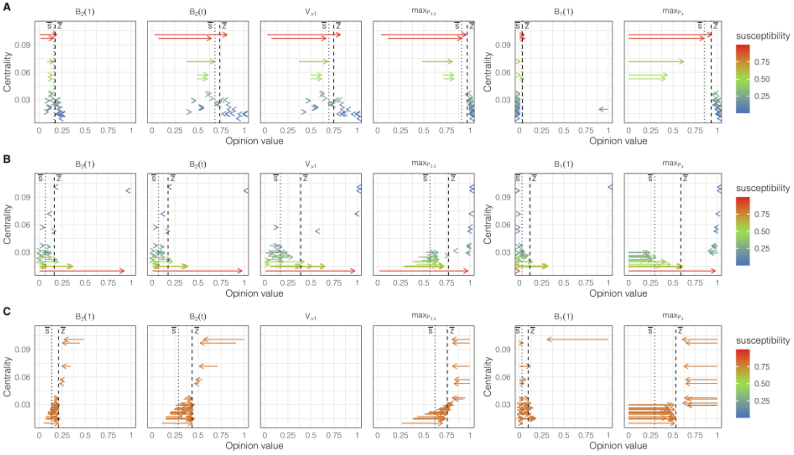
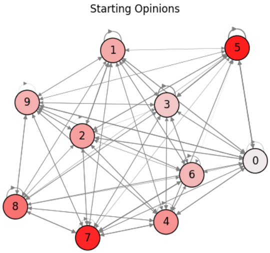


Table of Contents

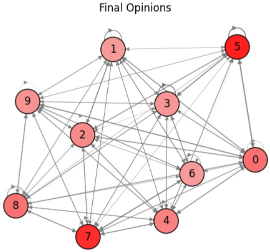
Reproduction of results with random graph

s	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-62.89	-9.19	-4.19	-0.29	-4.34	-4.68
$s_{B_2(1)}$	-2.05	-0.28	-0.14	0.01	0.14	0.16
$s_{B_2(t)}$	-2.05	-0.28	-0.14	0.01	0.14	0.16
$s_{max}^{P_2P_3}$	-1.74	-0.23	-0.12	0.01	0.15	0.15
$s_{B_1(1)}$	-13.51	-2.13	-0.90	-0.06	-0.88	0.17
$s_{max}^{P_4}$	-53.81	-7.29	-3.59	-0.19	-2.81	0.70

Table: gFJ in random graph, reproduction of paper results.



(a) Random network, $s_{B_2(1)}$



(b) Random network, z_{final}

Karate Club Dataset

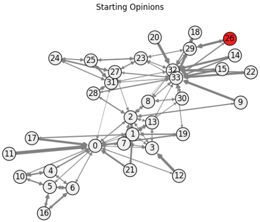
$\lambda = 0.8$, λ proportional to centrality, λ reverse proportional to centrality respectively.

	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-321.39	-14.71	-9.45	-0.26	-8.69	-8.96
$s_{B_2}(1)$	-53.42	-8.88	-1.57	0.07	2.52	5.46
$s_{B_2}(t)$	-53.42	-8.88	-1.57	0.07	2.52	5.46
$s_{max}^{P_2,P_3}$	-85.65	-7.08	-2.52	0.16	5.56	5.56
$s_{B_1}(1)$	-27.65	-6.48	-0.81	-0.02	-0.54	2.23
$s_{max}^{P_4}$	-252.99	-23.11	-7.44	0.04	1.42	9.10

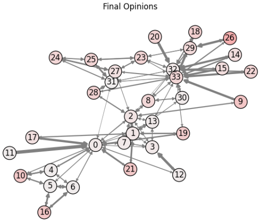
	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-96.98	-10.77	-2.85	-0.08	-2.84	-2.19
$s_{B_2}(1)$	-1.58	-0.21	-0.05	0.00	0.05	0.05
$s_{B_2}(t)$	-1.58	-0.21	-0.05	0.00	0.05	0.05
$s_{max}^{P_2,P_3}$	-1.49	-0.22	-0.04	0.00	0.05	0.05
$s_{B_1}(1)$	-9.35	-0.52	-0.28	-0.01	-0.27	0.04
$s_{max}^{P_4}$	-75.81	-9.10	-2.23	-0.05	-1.87	0.29

	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-125.52	-6.87	-3.69	-0.11	-3.68	-2.48
$s_{B_2}(1)$	0.84	-1.70	0.02	0.02	0.59	1.54
$s_{B_2}(t)$	0.84	-1.70	0.02	0.02	0.59	1.54
$s_{max}^{P_2,P_3}$	-26.93	-2.40	-0.79	0.06	2.01	2.01
$s_{B_1}(1)$	-0.81	-1.37	-0.02	0.00	0.07	1.07
$s_{max}^{P_4}$	-85.82	-9.11	-2.52	0.01	0.31	3.74

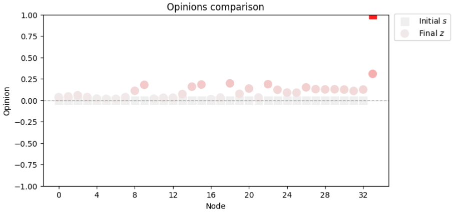
Karate Club dataset plots for $\lambda = 0.8$ and $s_{B_1(1)}$



(a) Karate club network, $s_{B_1(1)}$

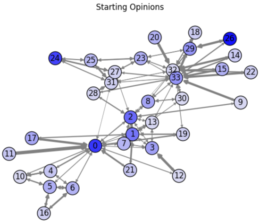


(b) Karate club network, z_{final}

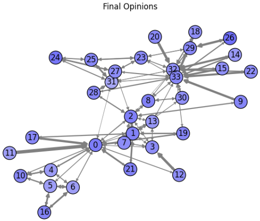


(c) Karate club network, initial and final opinion vectors.

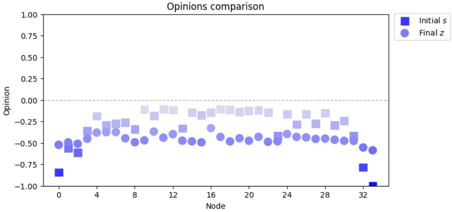
Karate Club dataset plots for $\lambda = 0.8$ and $s_{B_2(1)}$



(a) Karate club network, $s_{B_2(1)}$



(b) Karate club network, z_{final}



(c) Karate club network, initial and final opinion vectors.

Barabasi Albert network

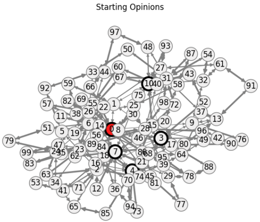
$\lambda = 0.8$, λ proportional to centrality, λ reverse proportional to centrality respectively.

	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-2730.45	-46.01	-27.30	-0.27	-27.32	-32.37
$s_{B_2}(1)$	-258.03	-14.92	-2.58	0.04	4.03	11.60
$s_{B_2}(t)$	-258.03	-14.92	-2.58	0.04	4.03	11.60
$s_{max}^{P_2,P_3}$	-426.52	-15.74	-4.27	0.12	11.98	11.98
$s_{B_1}(1)$	-83.35	-7.06	-0.83	-0.01	-0.67	3.18
$s_{max}^{P_4}$	-2097.50	-68.86	-20.97	-0.01	-1.46	22.54

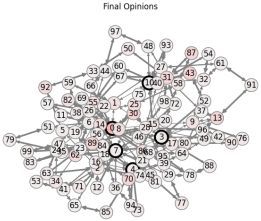
	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-2097.50	-68.86	-20.97	-0.01	-1.46	22.54
$s_{B_2}(1)$	-1.86	-0.11	-0.02	0.00	0.02	0.02
$s_{B_2}(t)$	-1.86	-0.11	-0.02	0.00	0.02	0.02
$s_{max}^{P_2,P_3}$	-1.79	-0.11	-0.02	0.00	0.02	0.02
$s_{B_1}(1)$	-20.76	-0.55	-0.21	-0.00	-0.21	0.02
$s_{max}^{P_4}$	-468.57	-21.01	-4.69	-0.04	-4.41	0.24

	Δ_{GDI}	Δ_{NDI}	Δ_{P_1}	Δ_{P_2}	Δ_{P_3}	Δ_{P_4}
s_{unif}	-545.31	-12.65	-5.45	-0.05	-5.41	-4.21
$s_{B_2}(1)$	0.51	-0.87	0.01	0.00	0.30	1.28
$s_{B_2}(t)$	0.51	-0.87	0.01	0.00	0.30	1.28
$s_{max}^{P_2,P_3}$	-96.59	-3.35	-0.97	0.03	2.62	2.62
$s_{B_1}(1)$	-1.58	-0.82	-0.02	0.00	0.00	0.66
$s_{max}^{P_4}$	-510.59	-16.82	-5.11	-0.01	-0.99	5.46

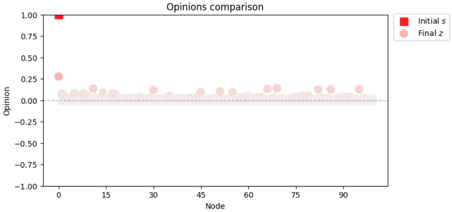
Barabassi Albert network plots with $\lambda = 0.8$ and $s_{B_1(1)}$



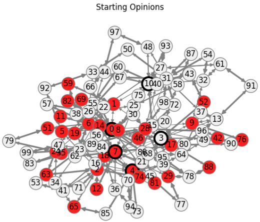
(a) Barabassi Albert network, $s_{B_1}(1)$



(b) Barabassi Albert network, z_{final}



(c) Barabassi Albert network, initial and final opinion vectors.



(a) Barabassi Albert network, $s_{B_2}(1)$

(b) Barabassi Albert network, z_{final}

(c) Barabassi Albert network, initial and final opinion vectors.

Table of Contents

Class-related Takeaway

- Class theoretical tools used in paper:
  - Graphs Notation
  - Social Networks: Notation, **Barabasi-Albert** model.
  - **PageRank centrality** used for susceptibility values for the paper results.
  - Occurrence of a non-Convex and a **linear optimization** problem (Theorems 5 and 6 respectively).
- Other theoretical tools:
  - **Eigenvalue decomposition** used in Theorem 5 for analytical construction of some polarizing initial vectors **s**.
  - Definitions from analysis (l_2-Ball, subspaces).
- Estimated difficulty:
  - Despite the complex and massive formulation, the core concepts and ideas are easy to understand and elegant.
  - Proofs of the theorems are a bit harder to understand (as usual) but don't prevent from understanding the points of the paper.

Thank you for your attention![1]

[1] Dynamics of opinion polarization, E. Biondi https://arxiv.org/abs/2206.06134