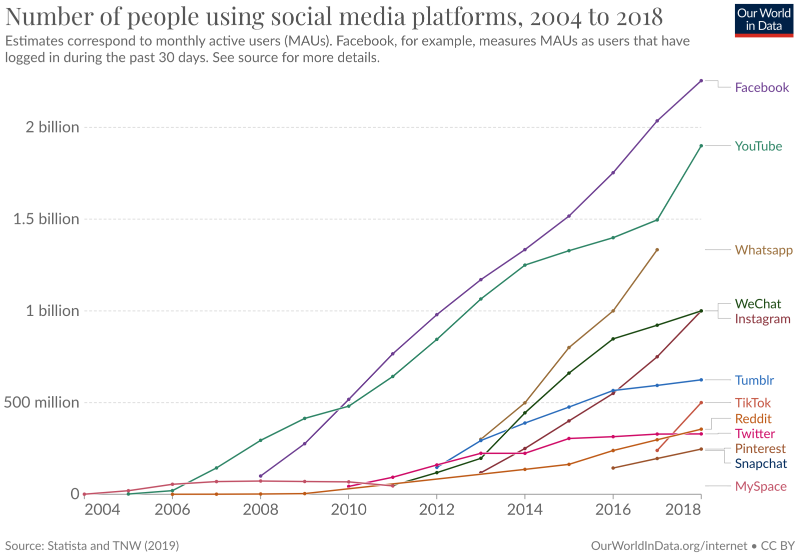
Impact of Censorship in Social Media

Kacper Krasowski

June 30, 2025

Outline

Introduction



Introduction

- **Research Question**
  - What are the implications of limiting freedom of speech on social media for overall welfare?

Introduction

- **Research Question**
  - What are the implications of limiting freedom of speech on social media for overall welfare?
- **Motivation**
  - Great and growing impact of social media on opinion dynamics
  - Increasing interest of regulators on social media (e.g. The Digital Services Act)
  - Existing gap in the literature

Introduction

- **Research Question**
  - What are the implications of limiting freedom of speech on social media for overall welfare?
- **Motivation**
  - Great and growing impact of social media on opinion dynamics
  - Increasing interest of regulators on social media (e.g. The Digital Services Act)
  - Existing gap in the literature
- **Contribution**
  - A new way of modeling the censorship in social media and examining its impact.
  - Filling the gap of quantitative analysis of freedom of speech in social media.

Introduction

- **The model**:
  - Agents' **opinions** are **repeated weighted averages** of their neighbours' opinions and their initial opinion.
  - **Censorship** introduced by **banning** agents with **extreme** enough **opinions**.
  - Welfare investigated in terms of:
    - **Polarization** - How much opinions differ in a network.
    - **Disagreement** - How much opinions differ among connected agents.
    - **Internal Conflict** - How much opinions have evolved.
    - **Mix** of the above indices.

Introduction

- **The model**:
  - Agents' **opinions** are **repeated weighted averages** of their neighbours' opinions and their initial opinion.
  - **Censorship** introduced by **banning** agents with **extreme** enough **opinions**.
  - Welfare investigated in terms of:
    - **Polarization** - How much opinions differ in a network.
    - **Disagreement** - How much opinions differ among connected agents.
    - **Internal Conflict** - How much opinions have evolved.
    - **Mix** of the above indices.
- **Preview of results**:
  - Higher **censorship increases polarization** and **reduces internal conflict**.
  - The impact on **disagreement depends** on the network size and connectivity.
  - There **exists** a **optimal censorship**.

Literature Review

- **Opinion Dynamics**
  - DeGroot (1974)
  - Friedkin and Johnsen (1997)
  - Golub and Jackson (2010)
  - Cameron Musco, Christopher Musco, and Tsourakakis (2018)

Literature Review

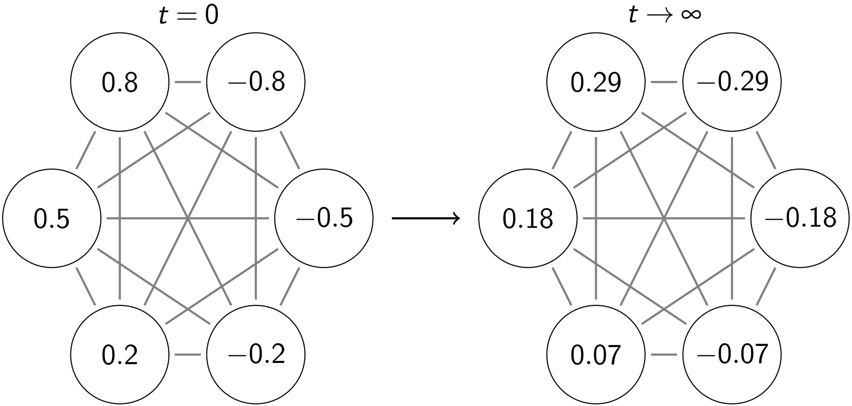
- **Opinion Dynamics**
  - DeGroot (1974)
  - Friedkin and Johnsen (1997)
  - Golub and Jackson (2010)
  - Cameron Musco, Christopher Musco, and Tsourakakis (2018)
- **Freedom of speech in social media**
  - Dehghan (2018)
  - Segado-Boj and Campo Lozano (2020)

Literature Review

- **Opinion Dynamics**
  - DeGroot (1974)
  - Friedkin and Johnsen (1997)
  - Golub and Jackson (2010)
  - Cameron Musco, Christopher Musco, and Tsourakakis (2018)
- **Freedom of speech in social media**
  - Dehghan (2018)
  - Segado-Boj and Campo Lozano (2020)
- **Impact of Social Media on Welfare**
  - Allcott et al. (2020)

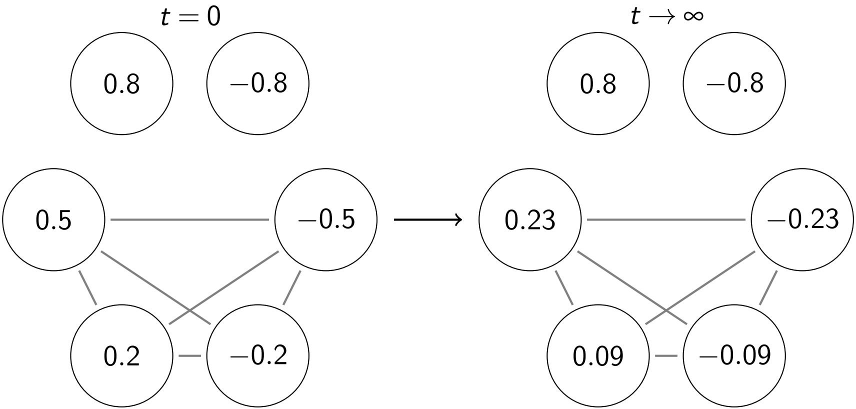
Example

Is climate change real?



Example

Is climate change real?



Set up

- Opinions $\bar{z}_i \in [-1, 1]$
  - $\bar{z}_i = 1$ - totally agree
  - $\bar{z}_i = 0$ - neutral
  - $\bar{z}_i = -1$ - totally disagree

Set up

- Opinions $\bar{z}_i \in [-1, 1]$
  - $\bar{z}_i = 1$ - totally agree
  - $\bar{z}_i = 0$ - neutral
  - $\bar{z}_i = -1$ - totally disagree
- Agents interact in a social network, represented by $G(V, E)$, where:
  - V is the set of nodes (agents)
  - E is set of edges (connections)

Set up

- Opinions $\bar{z}_i \in [-1, 1]$
  - $\bar{z}_i = 1$ - totally agree
  - $\bar{z}_i = 0$ - neutral
  - $\bar{z}_i = -1$ - totally disagree
- Agents interact in a social network, represented by $G(V, E)$, where:
  - V is the set of nodes (agents)
  - E is set of edges (connections)
- Two types of n agents:
  - Type 1 draws innate opinion $\bar{s}_i$ from distribution with mean $-\mu$ and variance σ^2.
  - Type 2 draws innate opinion $\bar{s}_i$ from distribution with mean μ and variance σ^2.

Set up

- Opinions $\bar{z}_i \in [-1, 1]$
  - $\bar{z}_i = 1$ - totally agree
  - $\bar{z}_i = 0$ - neutral
  - $\bar{z}_i = -1$ - totally disagree
- Agents interact in a social network, represented by $G(V, E)$, where:
  - V is the set of nodes (agents)
  - E is set of edges (connections)
- Two types of n agents:
  - Type 1 draws innate opinion $\bar{s}_i$ from distribution with mean $-\mu$ and variance σ^2.
  - Type 2 draws innate opinion $\bar{s}_i$ from distribution with mean μ and variance σ^2.
- Agents opinions evolve according to:

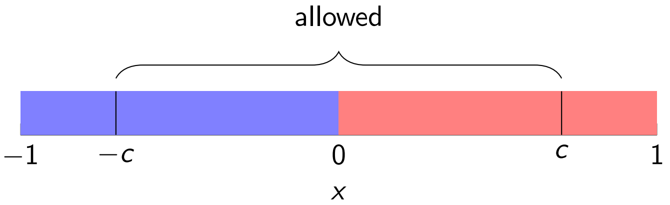
$$\bar{z}_i^{(t)} = \frac{\bar{s}_i + \sum_j a_{ij} \bar{z}_j^{(t-1)}}{1 + d_i}, \quad \bar{\mathbf{z}} = (\mathbf{I} + \mathbf{D} - \mathbf{A})^{-1} \bar{\mathbf{s}}$$

Censorship and Dynamics

- **Network Administrator** decides a threshold [-c, c] of allowed opinions.
  - $c = 1$ - full freedom of speech
  - $c = 0$ - full censorship

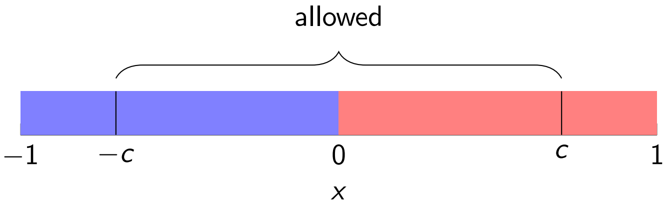
Censorship and Dynamics

- **Network Administrator** decides a threshold [-c, c] of allowed opinions.
  - $c = 1$ - full freedom of speech
  - $c = 0$ - full censorship



Censorship and Dynamics

- **Network Administrator** decides a threshold [-c, c] of allowed opinions.
  - $c = 1$ - full freedom of speech
  - $c = 0$ - full censorship



- The timings are as follows:
  1. Network Administrator learns the types, and intensity.
  2. Network Administrator decides the censor point.
  3. Opinions get drawn and evolve until they reach equilibrium.

Welfare Measures

Polarization - Variance of a set of opinions.

$$\mathscr{P} = \sum_{i=1}^{n} \bar{z}_i^2$$

Welfare Measures

Polarization - Variance of a set of opinions.

$$\mathscr{P} = \sum_{i=1}^{n} \bar{z}_i^2$$

Disagreement - How much opinions differ in the network.

$$\mathscr{D} = \sum_{i=1}^{n} \sum_{j \in N(i)} a_{ij} (\bar{z}_i - \bar{z}_j)^2$$

Welfare Measures

Polarization - Variance of a set of opinions.

$$\mathscr{P} = \sum_{i=1}^{n} \bar{z}_i^2$$

Disagreement - How much opinions differ in the network.

$$\mathscr{D} = \sum_{i=1}^{n} \sum_{j \in N(i)} a_{ij}(\bar{z}_i - \bar{z}_j)^2$$

Internal Conflict - How much opinions differ from the innate ones.

$$\mathscr{I}\mathscr{C} = \sum_{i=1}^{n} (\bar{z}_i - \bar{s}_i)^2$$

Welfare Measures

Polarization - Variance of a set of opinions.

$$\mathscr{P} = \sum_{i=1}^{n} \bar{z}_i^2$$

Disagreement - How much opinions differ in the network.

$$\mathscr{D} = \sum_{i=1}^{n} \sum_{j \in N(i)} a_{ij} (\bar{z}_i - \bar{z}_j)^2$$

Internal Conflict - How much opinions differ from the innate ones.

$$\mathscr{IC} = \sum_{i=1}^{n} (\bar{z}_i - \bar{s}_i)^2$$

Objective - Convex combination of the above indices.

$$\mathscr{W} = -[\alpha \mathscr{P} + \beta \mathscr{P} + (1 - \alpha - \beta) \mathscr{IC}]$$

Methodology

- Each agent talk with one another with intensity p.
- The point of interest is the perspective of the Network administrator.
- As the network administrator does not know which agents have what opinions, while deciding on c, from hers perspective each agent has a probability $F(c)$ to stay in the social network.
- Then the equilibrium opinions are given by:

$$\bar{\mathbf{z}} = (\mathbf{I} + \mathbf{D} - \mathbf{A})^{-1}\bar{\mathbf{s}} = (\mathbf{I} + F(c)np\mathbf{I} - F(c)p\mathbf{J})^{-1}\bar{\mathbf{s}}$$

Results

$$\mathbb{E}[\mathscr{P}] = \sigma^2 + \frac{1}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{D}] = \frac{F(c)np}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{I}\mathscr{C}] = \frac{(F(c)np)^2}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

Results

$$\mathbb{E}[\mathscr{P}] = \sigma^2 + \frac{1}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{D}] = \frac{F(c)np}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{I}\mathscr{C}] = \frac{(F(c)np)^2}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

1. **Polarization** is **increasing** in censorship

Results

$$\mathbb{E}[\mathscr{P}] = \sigma^2 + \frac{1}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{D}] = \frac{F(c)np}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{I}\mathscr{C}] = \frac{(F(c)np)^2}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

1. **Polarization** is **increasing** in censorship
2. **Disagreement** is:
   - **decreasing** in censorship if $\frac{1}{np} < 1$
   - initially **increasing** and then **decreasing** otherwise

Results

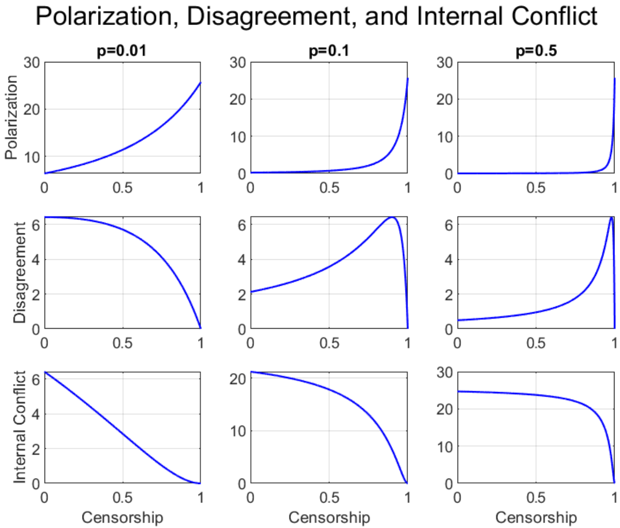
$$\mathbb{E}[\mathscr{P}] = \sigma^2 + \frac{1}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{D}] = \frac{F(c)np}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

$$\mathbb{E}[\mathscr{I}\mathscr{C}] = \frac{(F(c)np)^2}{(1+F(c)np)^2}(n\mu^2 + (n-1)\sigma^2)$$

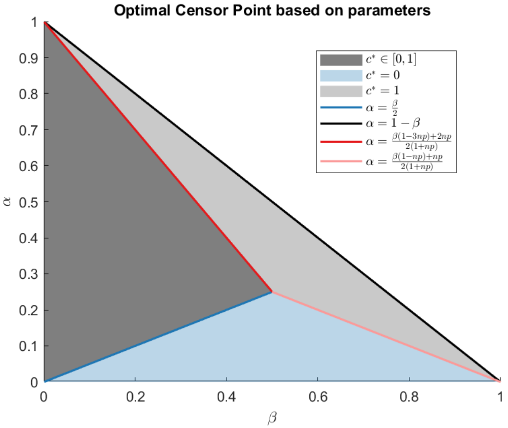
1. **Polarization** is **increasing** in censorship
2. **Disagreement** is:
   - **decreasing** in censorship if $\frac{1}{np} < 1$
   - initially **increasing** and then **decreasing** otherwise
3. **Internal conflict** is **decreasing** in censorship

Results

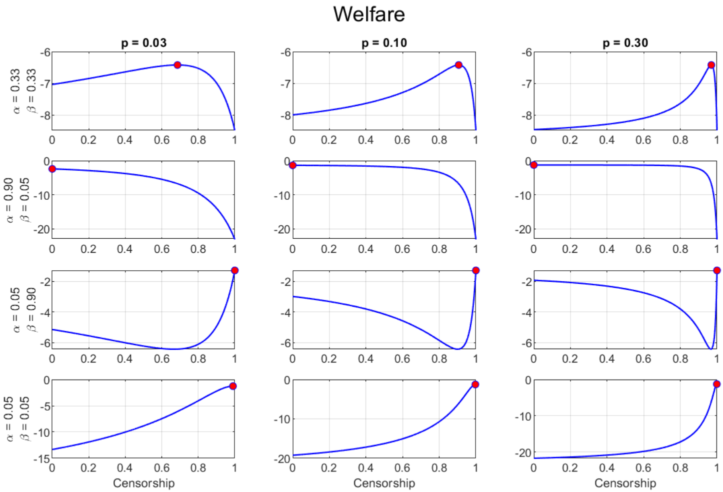

Polarization, Disagreement, and Internal Conflict
p=0.01
p=0.1
p=0.5
Polarization
Disagreement
Internal Conflict
Censorship
Censorship
Censorship
0
0.5
1
30
20
10
6
4
2

Results

If the none of the indexes have too much weight on it in the welfare function, then $c^* = Q(\frac{1}{np}\frac{2\alpha-\beta}{2-2\alpha-3\beta})$ is the optimal censoring point. Otherwise its either 0 or 1.



Results


Welfare
p = 0.03
p = 0.10
p = 0.30
α = 0.33 β = 0.33
α = 0.90 β = 0.05
α = 0.05 β = 0.90
α = 0.05 β = 0.05
Censorship
Censorship
Censorship

Extensions

No cross types communication

- **Polarization** is **increasing** in censorship
- **Disagreement** is:
  - **decreasing** in censorship if $\frac{2}{np} < 1$
  - initially **increasing** and then **decreasing** otherwise
- **Internal conflict** is decreasing in **censorship**
- $c^* = Q(\frac{2}{np}\frac{2\alpha-\beta}{2-2\alpha-3\beta})$
- Minor changes in indices caused by censorship

Extensions

No cross types communication

- **Polarization** is **increasing** in censorship
- **Disagreement** is:
  - **decreasing** in censorship if $\frac{2}{np} < 1$
  - initially **increasing** and then **decreasing** otherwise
- **Internal conflict** is decreasing in **censorship**
- $c^* = Q(\frac{2}{np}\frac{2\alpha-\beta}{2-2\alpha-3\beta})$
- Minor changes in indices caused by censorship

Different intensities across types

- **Polarization** is **increasing** in censorship
- **Internal conflict** is decreasing in **censorship**
- **Optimal censoring point** could be found by **numerical methods**.

Conclusions

- **Objective**: Investigate the effects of censorship in social media

Conclusions

- **Objective**: Investigate the effects of censorship in social media
- **Model**: Two types of agents with opposing opinions; censorship introduced by banning agents with opinions outside a set threshold. Opinions evolve according to Friedkin-Johansen model

Conclusions

- **Objective**: Investigate the effects of censorship in social media
- **Model**: Two types of agents with opposing opinions; censorship introduced by banning agents with opinions outside a set threshold. Opinions evolve according to Friedkin-Johansen model
- **Results**:
  - Higher censorship levels generally increase polarization and reduce internal conflict.
  - Disagreement initially rises with censorship but eventually decreases to zero.
  - Optimal censorship thresholds identified for specific parameters.

Avenue for Future work

- Investigate the incentives of network administrators, focusing on profit maximization rather than welfare.
- Incorporate censorship costs and user activity into profit functions.
- Explore empirical validation using data from platforms like Twitter or Reddit.

Thank You!

Appendix

$$\bar{z}_i^{(t)} = \frac{\bar{s}_i + \sum_j a_{ij}\bar{z}_j^{(t-1)}}{1+d_i} \tag{1}$$

$$\bar{\mathbf{z}}^{(\mathbf{t})} = (\mathbf{I}+\mathbf{D})^{-1}(\bar{\mathbf{s}} + \mathbf{A}\bar{\mathbf{z}}^{(\mathbf{t-1})}) \tag{2}$$

$$\bar{\mathbf{z}} = (\mathbf{I}+\mathbf{D})^{-1}(\bar{\mathbf{s}} + \mathbf{A}\bar{\mathbf{z}}) \tag{3}$$

$$(\mathbf{I}+\mathbf{D})\bar{\mathbf{z}} = (\bar{\mathbf{s}} + \mathbf{A}\bar{\mathbf{z}}) \tag{4}$$

$$(\mathbf{I}+\mathbf{D}-\mathbf{A})\bar{\mathbf{z}} = \bar{\mathbf{s}} \tag{5}$$

$$\bar{\mathbf{z}} = (\mathbf{I}+\mathbf{D}-\mathbf{A})^{-1}\bar{\mathbf{s}} \tag{6}$$

Appendix

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} p & \dots & p \\ \vdots & \ddots & \vdots \\ p & \dots & p \end{bmatrix} = p\mathbf{J}$$

$$\mathbf{D} = \begin{bmatrix} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_n \end{bmatrix} = \begin{bmatrix} \sum_i^n a_{1i} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_i^n a_{ni} \end{bmatrix} = \begin{bmatrix} np & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & np \end{bmatrix} = np\mathbf{I}$$

Appendix

$$\mathscr{P} = \sum_{i=1}^{n} \bar{z}_i^2 = \bar{\mathbf{z}}^{\mathbf{T}}\bar{\mathbf{z}} = \bar{\mathbf{s}}^{\mathbf{T}}(\mathbf{I}+\mathbf{L})^{-2}\bar{\mathbf{s}}$$

$$\begin{aligned}\mathscr{D} &= \sum_{i=1}^{n} \sum_{j\in N(i)} a_{ij}(\bar{z}_i - \bar{z}_j)^2 = \bar{\mathbf{z}}^{\mathbf{T}}\mathbf{L}\bar{\mathbf{z}} \\ &= \bar{\mathbf{s}}^{\mathbf{T}}(\mathbf{I}+\mathbf{L})^{-1}\mathbf{L}(\mathbf{I}+\mathbf{L})^{-1}\bar{\mathbf{s}}\end{aligned}$$

$$\mathscr{IC} = \sum_{i=1}^{n} (\bar{z}_i - \bar{s}_i)^2 = \bar{\mathbf{s}}^{\mathbf{T}}(\mathbf{I}+\mathbf{L})^{-1}\mathbf{L}^2(\mathbf{I}+\mathbf{L})^{-1}\bar{\mathbf{s}}$$

$$\begin{aligned}\mathscr{P} &= \bar{\mathbf{s}}^{\mathbf{T}} f_{\mathscr{P}}(\mathbf{L})\bar{\mathbf{s}}, \\ \mathscr{D} &= \bar{\mathbf{s}}^{\mathbf{T}} f_{\mathscr{D}}(\mathbf{L})\bar{\mathbf{s}}, \\ \mathscr{IC} &= \bar{\mathbf{s}}^{\mathbf{T}} f_{\mathscr{IC}}(\mathbf{L})\bar{\mathbf{s}}\end{aligned}$$

Appendix

$$
\begin{aligned}
\mathbb{E}[\mathscr{P}] &= \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathscr{P}}(\mathbf{L})\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{U}\Lambda_{\mathscr{P}}\mathbf{U}^{\mathbf{T}}\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}\Lambda_{\mathscr{P}}\bar{\mathbf{s}}_{\mathbf{U}}] \\
&= \sum_i^n \lambda_{\mathscr{P}i}\mathbb{E}[\bar{s}_{Ui}^2] = \sum_i^n \frac{1}{(1+\lambda_{Li})^2}\mathbb{E}[\bar{s}_{Ui}^2] \\
\mathbb{E}[\mathscr{D}] &= \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathscr{D}}(\mathbf{L})\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{U}\Lambda_{\mathscr{D}}\mathbf{U}^{\mathbf{T}}\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}\Lambda_{\mathscr{D}}\bar{\mathbf{s}}_{\mathbf{U}}] \\
&= \sum_i^n \lambda_{\mathscr{D}i}\mathbb{E}[\bar{s}_{Ui}^2] = \sum_i^n \frac{\lambda_{Li}}{(1+\lambda_{Li})^2}\mathbb{E}[\bar{s}_{Ui}^2] \\
\mathbb{E}[\mathscr{I}\mathscr{C}] &= \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}} f_{\mathscr{I}\mathscr{C}}(\mathbf{L})\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}^{\mathbf{T}}\mathbf{U}\Lambda_{\mathscr{I}\mathscr{C}}\mathbf{U}^{\mathbf{T}}\bar{\mathbf{s}}] = \mathbb{E}[\bar{\mathbf{s}}_{\mathbf{U}}^{\mathbf{T}}\Lambda_{\mathscr{I}\mathscr{C}}\bar{\mathbf{s}}_{\mathbf{U}}] \\
&= \sum_i^n \lambda_{\mathscr{I}\mathscr{C}i}\mathbb{E}[\bar{s}_{Ui}^2] = \sum_i^n \frac{\lambda_{Li}^2}{(1+\lambda_{Li})^2}\mathbb{E}[\bar{s}_{Ui}^2]
\end{aligned}
$$

Appendix

It is common knowledge in the SBM literature that the eigenvalues of $\mathbf{A}$ are:

$$\lambda_{A1} = F(c)n\frac{p+q}{2}, \lambda_{A2} = F(c)n\frac{p-q}{2}, \lambda_{A3} = ...\lambda_{An} = 0$$

And the eigenvalues of $\mathbf{D}$ are:

$$\lambda_{D1} = \lambda_{D2} = ...\lambda_{Dn} = F(c)n\frac{p+q}{2}$$

, then the eigenvalues of $\mathbf{L}$ are:

$$\lambda_{L1} = 0, \lambda_{L2} = F(c)qn, \lambda_{L3} = ... = \lambda_{Ln} = F(c)n\frac{p+q}{2}$$

Appendix

And from Chen, Lijffijt, and De Bie (2018) we know the mappings from eigenvalues of **L** to eigenvalues of each index:

1. $\lambda_{\mathscr{P}} = \frac{1}{(1+\lambda_L)^2}$
2. $\lambda_{\mathscr{D}} = \frac{\lambda_L}{(1+\lambda_L)^2}$
3. $\lambda_{\mathscr{I}\mathscr{C}} = \frac{\lambda_L^2}{(1+\lambda_L)^2}$

,thus:

$$\lambda_{\mathscr{P}1} = 1, \lambda_{\mathscr{P}2} = \frac{1}{(1+F(c)qn)^2}, \lambda_{\mathscr{P}3} = \ldots = \lambda_{\mathscr{P}n} = \frac{1}{(1+F(c)\frac{p+q}{2}n)^2},$$

$$\lambda_{\mathscr{D}1} = 0, \lambda_{\mathscr{D}2} = \frac{F(c)qn}{(1+F(c)qn)^2}, \lambda_{\mathscr{D}3} = \ldots = \lambda_{\mathscr{D}n} = \frac{F(c)\frac{p+q}{2}n}{(1+F(c)\frac{p+q}{2}n)^2},$$

$$\lambda_{\mathscr{I}\mathscr{C}1} = 0, \lambda_{\mathscr{I}\mathscr{C}2} = \frac{(F(c)qn)^2}{(1+F(c)qn)^2}, \lambda_{\mathscr{I}\mathscr{C}3} = \ldots = \lambda_{\mathscr{I}\mathscr{C}n} = \frac{(F(c)\frac{p+q}{2}n)^2}{(1+F(c)\frac{p+q}{2}n)^2},$$

Appendix

$$\mathbb{E}[\mathscr{P}] = \sigma^2 + \frac{1}{(1+F(c)nq)^2}(n\mu^2+\sigma^2)+(n-2)\frac{1}{(1+F(c)n^{\frac{p+q}{2}})^2}\sigma^2$$

$$\mathbb{E}[\mathscr{D}] = \frac{F(c)nq}{(1+F(c)nq)^2}(n\mu^2+\sigma^2)+(n-2)\frac{F(c)n^{\frac{p+q}{2}}}{(1+F(c)n^{\frac{p+q}{2}})^2}\sigma^2$$

$$\mathbb{E}[\mathscr{I}\mathscr{C}] = \frac{(F(c)nq)^2}{(1+F(c)nq)^2}(n\mu^2+\sigma^2)+(n-2)\frac{(F(c)n^{\frac{p+q}{2}})^2}{(1+F(c)n^{\frac{p+q}{2}})^2}\sigma^2$$