9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Network Segregation Bubble Size

Keywords: network analysis, segregation, random walk, page rank, segregation bubble

Extended Abstract

Introduction Persons in social networks often tend to have ego networks with persons that are similar to themselves. In context of a real world population network this mechanism leads to segregation: persons are more often connected to their own social group and less connected to persons from a different group. More then might be expected from population composition. Segregation is an important phenomenon affecting social cohesion and societal polarization.

Ballester and Vorsatz (2014) and van der Laan et al. (2022) use a random walk on a social network to calculate exposure and segregation scores for different groups. The use of random walks in network science is ubiquitous and has many forms (Masuda et al., 2017), but in this case a PageRank method (Page et al., 1998) is used: starting from the ego, the procedure follows links to connected persons after which there is a probability of $\alpha = 0.85$ of continuing the random walk. The probability of finishing the random walk on a person is the importance of that person to the ego. The importance weighted frequency of group labels in the neighborhood of the ego measures the exposure of the ego to the groups. The validity of value 0.85 in PageRank is unclear to the authors, but is often taken as a given. It may be an optimal value for its original use, ranking web pages, but social and population networks have a different structure and mechanism. In this presentation we examine and present educational exposure and segregation scores as a function of α using a population network with 18M persons, leading to interesting insights. Furthermore we look into the contributions induced by α to the exposure scores at the different network distances.

Data We use a whole population network which is derived by Statistics Netherlands using official administrative government registers (van der Laan et al., 2022). It includes family members, household members, colleagues, class mates and neighbors at multiple times points, from January 1st 2009 yearly up until January 1st 2020. For persons between 25 to 55 years old we have attained education in the 4 categories lower, middle, bachelor and master level, making it an interesting source for studying educational segregation.

Methods Using the random walk method of Ballester and Vorsatz (2014) we calculate for each inhabitant i in the Netherlands with education e the exposure $E_{il}(\alpha)$ to persons with an education l using different values for α, with e and $l \in \{\text{lower}, \text{middle}, \text{bachelor}, \text{master}\}$. α is the probability of continuing the random walk. Therefore, larger values of α lead to a higher contribution of persons further removed from the ego. In order to measure this we calculated for a sample of egos the contribution $C_{id}(\alpha)$ of persons in the network as a function of distance d (shortest path).

Results Figure 1 shows the average exposure $\langle E_{el}(\alpha)\rangle$ of persons with education e to education l as a function of α. For each group e the average self-exposure $\langle E_{ee}\rangle$ first increases with α and then decreases towards the population fraction (dashed line) of l. So for each $\langle E_{ee}\rangle$ there is a value for α at which it is maximal, suggesting that there is an optimal bubble size

for this type of segregation. Furthermore the figure shows the added value of the random walk method: restricting a segregation analysis to direct neighbors assumes bubble size 1, while a random walk allows for determining a bubble size. Interestingly, the maximum self-exposure for each education level e occurs at a (slightly) different value for α. To optimize for educational segregation we settled for $\alpha = 0.4$ instead of the default value of 0.85 (vertical dashed line)

It is helpful to see how much the persons at network distance d contribute to the exposure scores of a person with education e. Choosing a value for α induces a effective random walk length. $\alpha = 1$ means that the whole network contributes equally to the each exposure score. Figure 2 shows the average contribution $\langle C_{ed}(\alpha)\rangle$ in exposure (and segregation) scores for education levels e as a function of shortest-path distance. For $\alpha = 0.4$ the total contribution of the network decreases for larger distances from the ego, while for $\alpha = 0.85$ all distances contribute approximately equally. It seems more plausible that persons closer to the ego play a more important role in the exposure (and segregation) score of the ego. Therefore, a value of 0.4 seems more plausible than a value of 0.85.

Conclusion To calculate exposure and segregation scores on a network, a PageRank random walk method adds value to an analysis only using direct network neighbors. It is worthwhile to tune the α parameter, since it reveals an optimal bubble size for the segregation at interest.

References

C. Ballester and M. Vorsatz. Random walk-based segregation measures. *Review of Economics and Statistics*, 96:383–401, 2014. doi: 10.1162/REST_a_00399.

N. Masuda, M. A. Porter, and R. Lambiotte. Random walks and diffusion on networks. *Physics Reports*, 716-717:1–58, 2017. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2017.07.007.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

D. van der Laan, E. de Jonge, M. Das, S. Te Riele, and T. Emery. A whole population network and its application for the social sciences. *European Sociological Review*, jcac026, 2022. doi: 10.1093/esr/jcac026.

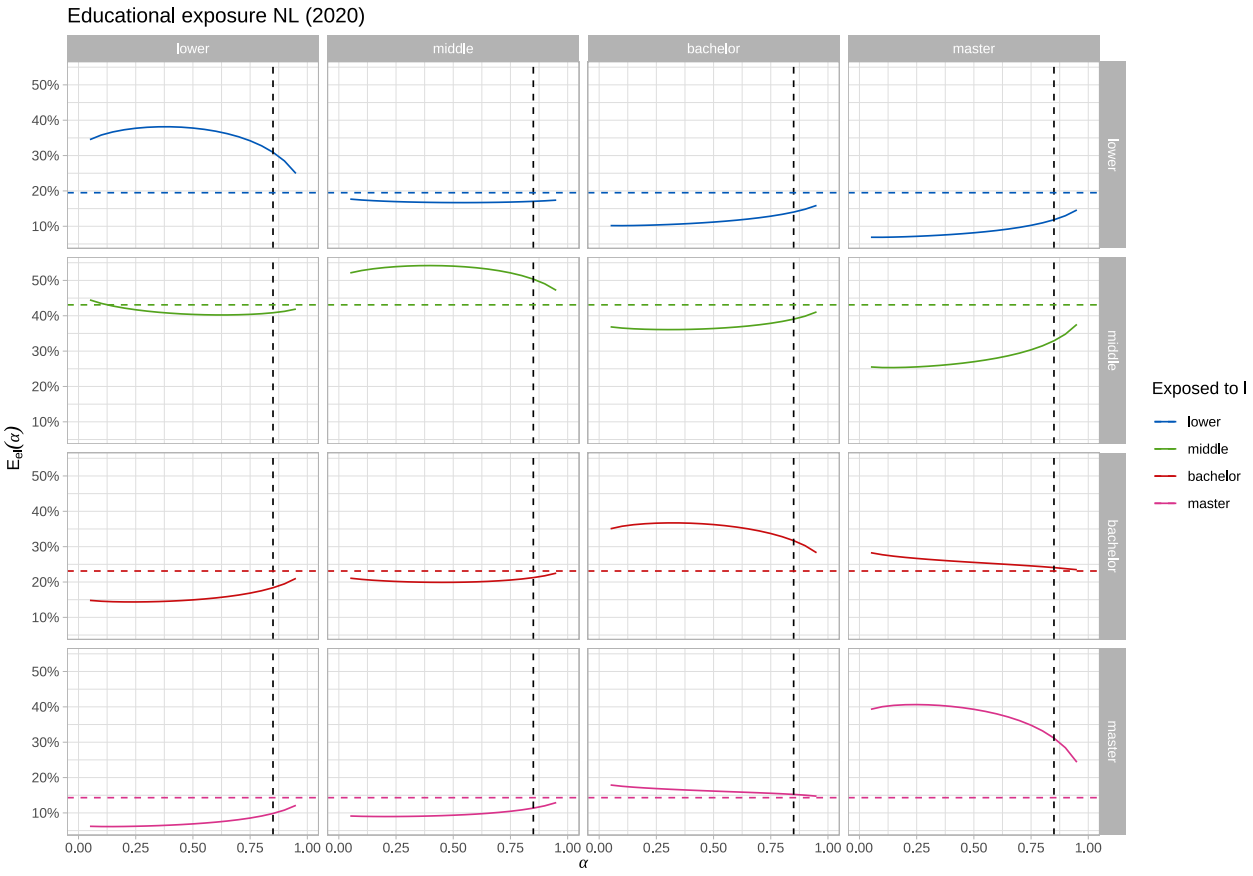

Figure 1: Average exposure $\langle E_{el}(\alpha)\rangle$ for education level e to education level l for the different education levels as a function of α. Colored dashed lines are the population fractions for l, vertical dashed line is default value $\alpha = 0.85$

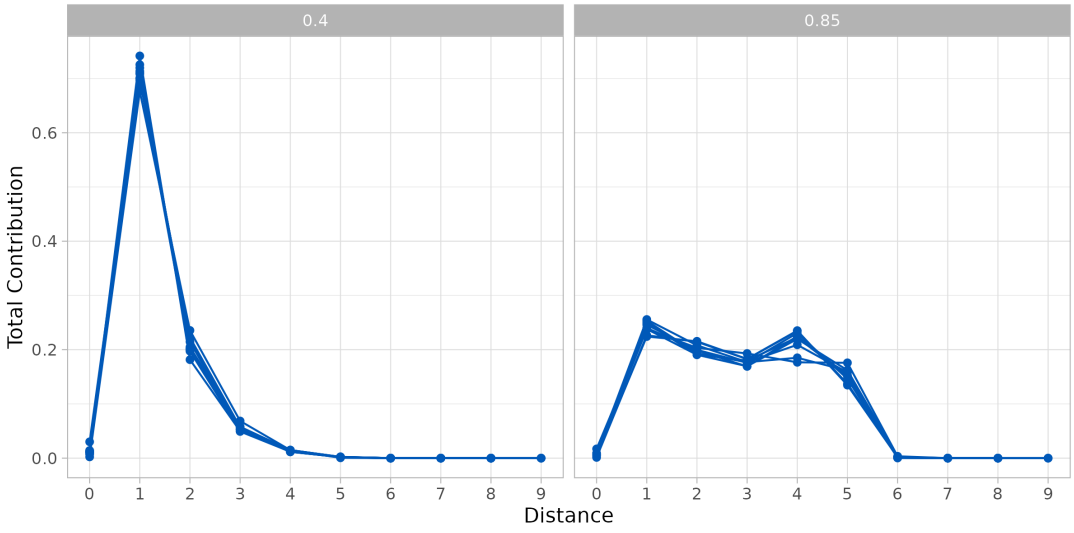

Figure 2: Contribution of network distances for $\alpha = 0.4$ and $\alpha = 0.85$