Journal of Computational Social Science (2023) 6:877–921
https://doi.org/10.1007/s42001-023-00216-9

RESEARCH ARTICLE

Ethnic segregation and spatial patterns of attitudes: studying the link using register data and social simulation

Thomas Feliciani[1] · **Jochem Tolsma**[1,2] · **Andreas Flache**[1]

Received: 26 September 2022 / Accepted: 12 June 2023 / Published online: 30 June 2023

Abstract

We theorize the causal link between ethnic residential segregation and polarization of ethnic attitudes within and between ethnic groups (e.g. attitudes towards immigration policies, multiculturalism, tolerance or trust in certain ethnic groups). We propose that the complex relationship between segregation and polarization might be explained by three assumptions: (1) ethnic membership moderates social influence–residents influence each other's attitudes and their ethnic background moderates this influence; (2) spatial proximity between residents increases opportunities for influence; (3) the degree of ethnic segregation varies across space–and therefore, the mix of intra- and inter-ethnic influence also varies across space. We borrow and extend an (agent-based) simulation model of social influence to systematically explore how these three assumptions affect the polarization of ethnic attitudes within and between ethnic groups under the assumptions made in the model. We simulate neighborly interactions and social influence dynamics in the districts of Rotterdam, using empirically observed segregation patterns as input of our simulations. According to our model, polarization in ethnic attitudes is stronger in districts and parts of districts where mixing of ethnic groups allows for many opportunities to interact with both the ethnic ingroup and the outgroup. Our study provides a new theoretical perspective on polarization of ethnic attitudes by demonstrating that the segregation-polarization link can emerge as an unintended outcome from repeated intra- and inter-ethnic interactions in segregated spaces.

Keywords Segregation · Polarization · Social influence · Agent-based modeling

✉ Thomas Feliciani
t.feliciani@rug.nl

[1] ICS/Department of Sociology, University of Groningen, Groningen, The Netherlands

[2] ICS/Department of Sociology, Radboud University, Nijmegen, The Netherlands

Introduction

Ongoing immigration flows contribute to the diverse ethnic composition of many western countries. Concurrently, we observe increasing ethnic residential segregation in many destination countries between and within cities [1–3]. This has impacted our societies greatly. Many are concerned that ethnic diversity and segregation may erode cohesion and lead to extremization and polarization in ethnic attitudes, such as attitudes towards immigration or about the trustworthiness of certain ethnic groups. Arguments put forward in the literature for the presumed links between the ethnic composition of the residential environment and the distribution of specific ethnic attitudes often refer to conflict or anomie mechanisms. Ethnic diversity, density and/or segregation would increase feelings of ethnic economic and cultural threat, unsafety and anomie among all residents of these areas. Because these feelings, in turn, are important determinants of many ethnic attitudes these mechanisms at the neighborhood or context-level would explain why attitudes such as distrust in ethnic minorities or outgroups are more prevalent in some geographic areas than others [4].

In this paper we argue that the literature on neighborhood effects does not draw a complete picture of how ethnic segregation between and within neighborhoods affects ethnic attitudes. For one, the empirical evidence is mixed: ethnic diversity is not consistently related to outcomes like a deterioration of social cohesion between and within ethnic groups; and outgroup sizes at the local level are not consistently related to more feelings of ethnic group threat [4]. Secondly, the causal link between ethnic residential patterns and the development of ethnic attitudes might be undertheorized. We claim that macro-level neighborhood effects, such as proposed in the threat or anomie mechanisms, are not necessary for the emergence of extremization and polarization of ethnic attitudes. This link might also emerge from a small set of psychologically motivated assumptions about influence processes taking place at the micro-level of neighborly interactions which so far have been underused in the literature. Ethnic segregation affects opportunities for intra- and inter-ethnic interaction. In turn, inter- and intra-ethnic interactions and the resulting micro-level processes of attitude change between contact partners may very well impact whether and how ethnic attitudes become extreme or polarized. Thus, observed macro-level relationships between ethnic residential patterns and the spatial distribution of ethnic attitudes might emerge from micro-level processes of attitude change among neighbors alone.

Our proposed "social influence" mechanism hinges on complex neighborly interactions between many individuals over long periods of time and with possible non-linear effects on their attitude changes. It is therefore hard to predict, based on intuition alone, which patterns of segregation are–according to this mechanism–more likely to lead to polarization in ethnic attitudes. Therefore, we elaborate our assumptions about social influence in an agent-based model (ABM)[1] and

[1] ABMs are formal and computational models used to simulate social systems. They are used to study 'in silico' complex emergent phenomena (such as residential segregation and attitude polarization) by simulating, under controlled conditions, repeated interactions among individuals (the agents) that follow

perform computational experiments to assess their implications for the relation between segregation and polarization in ethnic attitudes. With the ABM we intend to show that micro-level processes of attitude change alone–under the right conditions and without assuming macro-level neighborhood effects–can give rise to a complex relation between segregation and polarization.

Several ABMs have been used to investigate the relationship between residential segregation and intra- and inter-group dynamics (e.g. [5, 6]) and intra- and inter-group social influence [7–9]. Various behavioral theories of interpersonal dynamics inspire these ABMs and can be relevant here. We chose to demonstrate our proposed social influence mechanism by extending one of these ABMs–the "repulsive influence" model (*RI-model* for short) [8, 9], based on the theories of balance and cognitive dissonance [10, 11] and the repulsion hypothesis [12]. In a nutshell, individuals in the RI-model adjust their ethnic attitudes in response to the exposure to the attitudes of others they interact with. These attitude changes are assumed to be in the direction of minimizing the cognitive dissonance that arises, e.g., from disagreeing with an ingroup neighbor or agreeing with an outgroup member on an ethnically-salient topic. Depending on the extent of the disagreement and on the ethnic membership of the individuals, cognitive dissonance can be resolved by either reducing the attitude difference with the interaction partner, or by amplifying it.

Previous work on the RI-model already suggested that the degree to which attitudes polarize within and between groups is linked to how the groups are arranged–or segregated–in the simulated environment [8, 9]. However, the relevance of these findings for real-world urban environment is an open question, as research on models of opinion dynamics (and specifically on the RI-model) hinged on highly artificial assumptions about ethnic residential patterns. Our work aims to fill this gap by exploring the attitude dynamics of the RI-model in more realistic simulated urban environments.

We achieve this by seeding the initial simulated residential environment with empirical data on the demographic composition of 12 administrative districts of Rotterdam, a Dutch city of almost 600,000 inhabitants.[2] We focus in particular on the spatial distribution within these districts of residents with and without a non-western migration background. Residents with a non-western migration background constitute 38.4% of all Rotterdam residents. The category 'non-western' encompasses mostly visible minority groups, predominantly residents with a Turkish, Moroccan or Surinamese ethnic descent. The umbrella terms 'western' and 'non-western' are commonly used in the Dutch migration debate and although they may be outdated [13], Statistics Netherlands still makes use of these labels for reasons of consistency. The seeding of the model with real segregation patterns responds to the call by the scholarly community for empirically rooted agent-based models [14–16], specifically for spatially explicit models [17] and models of attitude change [7, 18]. Empirical seeding of the input of spatial ABMs such as ours has examples in earlier

Footnote 1 (continued)

simple, empirically plausible or theoretically relevant behavioral rules. See [60, 61] for an introduction to agent-based modeling.

[2] Reference year: 2014, used throughout the study. Source: Statistics Netherlands, 2018.

related work [5, 6] and is an important step for reproducing empirically observed phenomena and for formulating empirically falsifiable hypotheses.

With our study we also improve on the methods for measuring the simulation outcomes: we adopt metrics of spatial correlation that allow for a better understanding of when and where the RI-model predicts ethnic attitudes to polarize and under which conditions. We particularly focus on the *alignment* of ethnic attitudes and ethnic membership. We speak of alignment when residents adopt attitudes similar to members of their ingroup and opposite to members of their outgroup [8, 9] –a pattern that can signal the deterioration of interethnic relationships.

In sum, with this contribution we aim to understand whether and how, according to the RI-model, empirically-observed ethnic residential segregation patterns affect the polarization of ethnic attitudes within and between ethnic groups. We hereby contribute to different strands of research: on neighborhood effects, on the diversity-cohesion link, on models of social influence, and on the empirical seeding of agent-based models. In the remainder of this article we define and operationalize the concept of ethnic residential segregation ("Interaction patterns in segregated localities"), we introduce the RI-model and the outcome variables ("The repulsive influence (RI) model"), describe our simulation experiment and present our ("Results"), and discuss our findings ("Conclusion").

Interaction patterns in segregated localities

Our attitudes can be influenced by others we interact with. For example, depending on the circumstances, we might be persuaded by others [19–21]; adjust our attitude in order to comply with a norm [22, 23]; imitate others [24]; or we might strive to minimize or maximize the difference between our attitudes and those of our social surroundings [11, 25]. Some social categories can play a role in these influence processes. When two individuals are forming an opinion about anti-immigration policies, for example, the discussion can have different outcomes depending on whether either of them is an immigrant.

While there are many forms of social interactions that influence ethnic attitudes, here we focus on neighborly interactions. We thus abstract from other forms of influence through, e.g. (social) media, hearsay about attitudes or behavior of member of other groups, or workplace encounters, and focus on the isolated impact of neighborly interactions. In our daily neighborly interactions we are exposed to some social categories more than we are to others. This is for two reasons. The first is that our local living environment is the locus of many of our daily interactions. Interaction intensity generally decays with distance: spatial proximity increases the chances of face-to-face interaction between individuals—even in a time of cheap, fast mobility and of ubiquitous social media [26–30]. The second reason is residential segregation, famously defined as "the degree to which two or more groups live separately from one another, in different parts of the urban environment" [31]. Our living environments are, to varying degrees, ethnically segregated. Ethnic residential segregation thus determines with whom we are likely to interact and by whom our attitudes can be influenced [32].

In sum, our ethnic attitudes can be shaped by local interactions in ethnically segregated residential areas. Our general expectation is therefore that, when different residential areas feature different ethnic residential segregation patterns, this may, because of social influence processes alone, result in differences in the distribution and polarization of ethnic attitudes between and within such localities. To gauge the validity of this reasoning we simulate local interactions in residential areas that vary in segregation patterns.

In the remainder of this Section we motivate our choice for the districts of Rotterdam as our case study and introduce some descriptive statistics that characterize these districts ("Rotterdam districts as case study"); we then operationalize our independent variable, ethnic residential segregation, and measure it in the Rotterdam districts ("The spatial distribution of ethnic groups within Rotterdam").

Rotterdam districts as case study

The residential districts ('wijken' in Dutch) of Rotterdam are an ideal case study for our simulation experiment. For one, these districts differ in their population densities, composition and levels of segregation; all while being similar in many other regards, e.g. sharing the same political and economic context.

Secondly, these districts can be regarded almost as 'islands' within the city. Because of their large areas and population size, all of them but the smallest (district Pernis) are an aggregate of several neighborhoods ('buurten' in Dutch)–and so all districts but Pernis have their own residential and commercial areas, multiple supermarkets, schools etc. Moreover, as partly visible in Fig. 1, Panel A, these districts are often separated by physical barriers which are hard to cross, such as highways, railways or waterways (e.g. the large river Maas). All this makes it plausible to assume that *neighborly* interactions occur predominantly within districts rather than between. This in turn allows us to simulate interactions in these districts while treating the districts as if they were independent from each other–a necessary simplification dictated by the computational requirements of our experiment, and which we will evaluate later in the article (see Appendix A, "Slope of the distance decay function").

Another simplification in our proof-of-concept simulation experiment is that there are only two ethnic groups.[3] Again, the ethnic composition of Rotterdam districts makes this simplification acceptable to some degree. We focus on the divide between residents with and without a non-western migration background which is very visible, salient, and politically relevant in the Netherlands. The social disadvantages of the former group and the cultural differences between the two are the reason why Statistics Netherlands differentiates between residents with and without native Dutch ethnic background (i.e. Dutch natives and 1st or 2nd generation migrants); and those with a migration background are further distinguished by country of origin: western

[3] Relaxing this assumption adds a layer of complexity that cannot be thoroughly addressed in this one article.

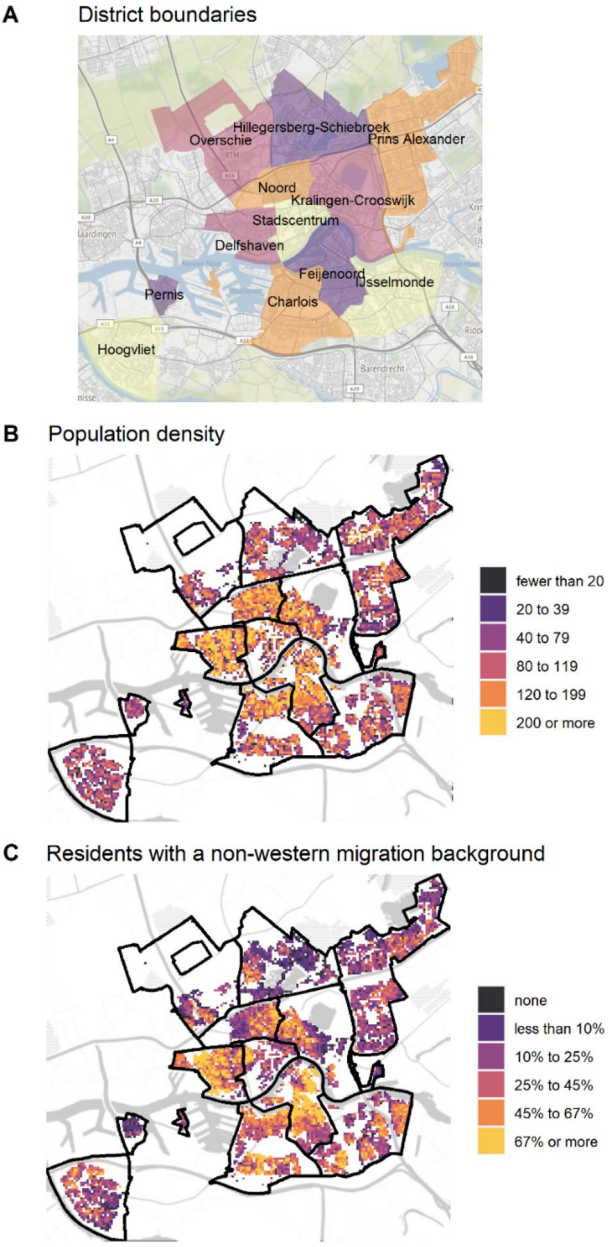

Fig. 1 Overview of the twelve chosen districts in Rotterdam: their administrative boundaries (Panel **A**); their population densities measured at a raster level (Panel **B**); and the percentage of the population with a non-western migration background–raster level (Panel **C**). Districts differ in area, population density and presence, size and degree of mixing of ethnic memberships (western/non-western descent). Credit: basemap imagery from Stamen 2023

or non-western ethnic background [33].[4] Its relevance and data availability make the ethnic divide between Dutch residents with and without a non-western migration background a fitting example for our study.

The municipality of Rotterdam consists of 22 administrative districts. Of these, 10 are industrial neighborhoods, uninhabited areas and a municipality exclave (Hoek van Holland). We focus on the 12 remaining districts, shown in Fig. 1. These range from the smallest district, Pernis—located in the south-west, with an area of 1,6 km^2 and a population of 4795—to the largest, Prins Alexander (north-east; area: 18,6 km^2; population: 93,920). Statistics Netherlands offers very fine-grained data on population densities and ethnic composition at the level of geographic areas spanning 100 by 100 m, which we label "square units" henceforward (see Fig. 1, panels B and C). The 12 districts of Rotterdam contain on average about 530 square units for which population densities were known.

The spatial distribution of ethnic groups within Rotterdam

In our theoretical model, the salient dimension of ethnic segregation is the degree to which individuals are exposed to the influence from members of their ethnic ingroup versus outgroup, reflecting the potential of inter- and intra-ethnic contact. We therefore characterize ethnic segregation as the degree of relative exposure to—or relative possibility of contact with—residents with/without a non-western ethnic background. We calculate each ethnic group's local outgroup exposure as the spatially weighed outgroup density [34]. This local outgroup exposure measure will depend on the size of the groups in the district, how groups are clustered in specific sub-areas of the district and on the evenness of the spatial distribution of the groups in the district and hence by extension also on the shape of the district.

Spatial indices of exposure are based on the notion of spatial proximity. With p_{ij} we denote the proximity between i and j, two of the N residents of the same district:

$$p_{ij} = \frac{\exp\left(\frac{-d_{ij}}{s}\right)}{\sum_{j \in N, j \neq i} \exp\left(\frac{-d_{ij}}{s}\right)} \tag{1}$$

where d_{ij} is the Euclidean distance "as the crow flies" between the two residents i and j, in meters; and $s \in \{10,100,1000\}$ is an arbitrary parameter which defines the slope of the distance decay function, allowing to measure segregation at different geographical scales. The denominator is a normalization constant ensuring that total proximity sums to one so that local outgroup exposure levels can be compared across districts. Based on the spatial proximity function we can define the local exposure E of resident i to the outgroup o by summing over all neighbors j who belong to the outgroup ($g_j = o$):

[4] Statistics Netherlands defines as "non-western" the whole of Africa and Latin America, plus Turkey and all Asian countries except for Indonesia and Japan. "Western" are all countries that are not "non-western".

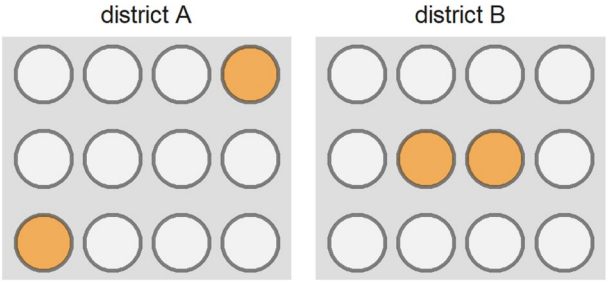

Fig. 2 Stylized districts with two ethnic groups: orange and white. The agents with the highest local outgroup exposure are the orange agents in district **A**; however, the average local outgroup exposure is higher in district **B**

$$E_{o,i} = \sum_{j}^{g_j=o} p_{ij} \tag{2}$$

Local outgroup exposure ranges from 0 to 1, where higher values signify stronger relative exposure of an individual to the other ethnic group (relative to the exposure to any agent). A score of 1 indicates that i is only exposed to ethnic out-group members (and hence only interacts with out-group members); and with a score of 0 i is only exposed to ethnic ingroup members.

We rely on the *average* local outgroup exposure to characterize districts. A district where the average outgroup exposure is low is one where members of the two groups will mainly mingle within their own groups and not between groups. For a district D, the average local outgroup exposure is:

$$E_{o,D} = \frac{\sum_i E_{o,i}}{N_D} \tag{3}$$

With evenly distributed groups, local outgroup exposure and average local outgroup exposure will be maximum when the two groups are of the same size. One might expect that districts with relatively high average local outgroup exposure would house more residents who are extremely exposed to outgroup neighbors, say an $E_{o,i} > 0.70$. However, this is not necessarily the case, as illustrated by two stylized districts in Fig. 2. In district A, two minority residents (orange) lie at opposite corners of the district, as far away from one another as possible. Here, the minority residents are highly exposed to the majority residents (white), and the majority residents are scarcely exposed to the minority. Thus, in district A the local outgroup exposure is on average relatively low, but there are two peaks of very high levels of local outgroup exposure. The situation is reversed in district B, where the two minority residents lie at the center of the district, close to one another, and all majority residents are near one or both. Thus, compared to district A, in district B there are no peaks of extremely high local outgroup exposure, while local outgroup exposure is higher on average.

Table 1 Descriptive statistics for the modelled districts in Rotterdam, assuming (s=100)

District name	Number of square units	Population	% Non-western	Average outgroup exposure of natives and Western	Average outgroup exposure of non-Western	Average outgroup exposure–all residents (s = 100)	% residents with local outgroup exposure > 0.7
Pernis	56	3905	9.01	0.085	0.877	0.155	8.81
Hillegersberg-Schiebroek	476	38,455	16.49	0.145	0.713	0.238	9.43
Prins Alexander	1025	91,215	22.69	0.206	0.748	0.322	14.34
Hoogvliet	431	32,825	25.23	0.220	0.698	0.333	12.4
Overschie	181	14,405	26.34	0.244	0.645	0.350	7.28
Stadscentrum	246	34,075	36.89	0.286	0.561	0.375	12.54
Noord	317	49,400	38.73	0.317	0.539	0.398	6.37
Kralingen-Crooswijk	363	48,505	36.67	0.319	0.559	0.405	5.48
IJsselmonde	607	56,255	36.86	0.325	0.577	0.413	7.22
Feijenoord	508	71,295	57.72	0.462	0.391	0.424	9.42
Charlois	535	64,275	46.81	0.408	0.509	0.453	3.03
Delfshaven	422	74,145	58.24	0.505	0.416	0.457	2.88
All 12 districts	**5167**	**578,755**	**36.37**	**0.308**	**0.524**	**0.386**	**8.14**

Different distance decay functions for the exposure index (i.e. for s=10 and s=1000) produce an almost identical ordering of districts. Variables are calculated on the square units for which demographics were known. Source: Statistics Netherlands, 2018. Reference year: 2014

Districts are ordered by outgroup exposure; city-level statistics (bold text) result from aggregating across districts

This seems to be the case for the 12 districts of Rotterdam as well. Table 1 summarizes the district-level statistics of exposure for the 12 districts of Rotterdam, ordered by their average local outgroup exposure ("all residents"). Districts are shown to differ by populated area (number of square units), demographics (population size and density of non-western residents), and average local outgroup exposure. The table shows that the percentage of highly-exposed residents (e.g., for which local outgroup exposure > 0.7) is indeed not always higher in districts with higher *average* local outgroup exposure. We will take this fact into consideration while exploring the effect of local outgroup exposure on the simulation dynamics.

Furthermore, from Table 1 we learn that districts with a high proportion of non-western residents are generally also characterized by strong average local outgroup exposure. The association between minority group size and average exposure is however not perfect and this is because ethnic groups are not evenly distributed within the districts. A small example of this are the districts Noord and Kralingen-Crooswijk: the non-western minority is larger in the former (closer to 50%), whereas the average exposure to the outgroup is higher in the latter.

The repulsive influence (RI) model

Here we introduce the RI-model: the intuition behind the model and its theoretical underpinnings; the expectations on the model behavior that can be drawn from previous research on this model ("Expectations"); the scheduling of a simulation run ("Ingredients of the ABM: agents' attributes, selection of interac-tion partner, social influence"), the seeding of the model with empirical data and its formal implementation ("Initialization and empirical seeding", "Opinion dynamics"); and the operationalization of the dependent variables ("Outcome measures").

The RI-model defines how individuals adjust their ethnic attitudes by interacting with each other. Initially proposed as an extension to classical models of social influence [35–37], the RI-model and its different implementations [38–40] builds on the cognitive theories of balance and dissonance [10, 11]. Depending on their ethnic membership, disagreement between individuals can give rise to dissonance, and the RI-model incorporates two ways in which dissonance can be minimized: *assimilation* and *repulsion*. Assimilation refers to one's tendency to reduce cognitive dissonance by reducing the degree of disagreement with the interaction partners. At the macro level, assimilation fosters consensus. By contrast, with repulsion the cognitive dissonance is resolved by amplifying the initial disagreement: this polarizes the attitudes.

Two factors affect whether attitude changes follow assimilation or repulsion: the extent of disagreement and the ethnic group membership. Small disagreements between two individuals result in assimilation, and thus tend to be resolved by reducing their attitude difference. When disagreement is sufficiently large, however, attitudes change repulsively, which increases the attitude difference. Ethnic group membership affects how large disagreement must be for repulsion to occur: when individuals disagree with others from the same group, cognitive dissonance is easily resolved by reducing disagreement. Thus, attitude repulsion requires large disagreement between ingroup members. By contrast, agreement with the outgroup constitutes

a kind of dissonance that is more easily resolved by increasing disagreement. In other words, the conditions for attitude assimilation are more easily met among people with the same group membership, and the reverse is true for attitude repulsion.

It has been shown that the RI-model can generate (1) attitude extremization, where agents' attitudes become on average more extreme than they were at the start of the simulation; (2) attitude polarization – that is, the splitting of the population into two subsets with large attitude differences between them and small attitude differences within them; (3) alignment of attitudes and ethnic membership – that is, attitude polarization between ethnic groups [8, 9]. Previous research on the RI-model also shows a negative association between ethnic segregation and polarization. Intuitively, this is because segregation minimizes the opportunity for interactions with the outgroup – fewer interactions with the outgroup reduce the potential for attitude repulsion, and thus attitude extremization and polarization. These predictions, it is argued, clash with what is predicted by alternative social influence processes and defy the common-sense belief that ethnic residential segregation exacerbates hostile attitudes between groups. The relevance of the social and scientific problem we are addressing warrants further scrutiny of the robustness and validity of the RI-model predictions. So far, the RI-model was studied in highly abstract simulation environments. For instance, the simulated population was divided into two equally-sized groups. Second, the population density was constant in every region of the world, and thus there was no spatial clustering of agents. Third, previous work explored population sizes up to 6400 agents, an arguably insufficient size for the representation of realistic segregation patterns and geographical shapes of real neighborhoods. Fourth, the investigated artificial segregation patterns were generated with the Schelling-Sakoda segregation model, which tends to generate locally homogeneous ethnic clusters, where interethnic interaction is only possible on the boundaries between clusters while being effectively precluded elsewhere. This does not resemble the spatial distribution of ethnic groups within real-world cities, where ethnic clusters are rarely perfectly homogeneous.

In sum, the artificial ethnic residential segregation patterns used in earlier simulation studies with the RI-model do not reflect the more complex and less-extreme forms of segregation empirical research found within real neighborhoods [41]. Our study relaxes this constraint, which allows to assess whether previously-observed model dynamics generalize to simulations with more realistic populations and spatial structures.

Expectations

Literature on the RI-model [8, 9] allows us to lay out a first set of expectations concerning how ethnic attitudes should be affected by outgroup exposure (defined in 2.2, Eqs. 2 and 3).

For instance, in the RI-model, contact with the outgroup (compared to ingroup contact) carries greater chances of causing mutual repulsion, which increases disagreement by making attitudes more extreme. Our assumption is that interactions decay with distance and thus that proximal neighbors interact more often than

non-neighbors do. Therefore, we expect that residents who are more exposed to their outgroup (i.e., higher local outgroup exposure scores) may need fewer neighborly interactions to become extremists than those who are less exposed to the outgroup in their local environment.

Expectation 1a) Agents who are locally more exposed to outgroup agents develop extreme attitudes after fewer interactions.

Next, we ask whether, at the district level, the attitudes are polarized. Generally, emerging properties of a complex social system cannot be easily inferred from its constituent parts (e.g. [42, 43])–this makes it particularly difficult to scale expectation 1a to the district level. We do so naïvely and conjecture that, ceteris paribus, higher average outgroup exposure tends to foster attitude repulsion. By the end of the simulation run, i.e. after a set maximum number of interactions, we therefore expect that attitude polarization will be stronger in districts with higher average outgroup exposure.

Expectation 1b) Districts with higher levels of average local outgroup exposure develop a higher degree of attitude polarization.

Expectations 1a and 1b look at agent attitudes without exploring how these attitudes develop within and between the two ethnic groups. Next, we investigate whether outgroup exposure affects the *alignment* between attitude valence (or 'sign') and ethnic group membership. In studying alignment, it is important to distinguish between alignment at the district level and at the local level. Alignment at the district level captures the extent to which the average attitudes within the two ethnic groups in a district differ from each other. Attitude alignment at the local level refers to the extent to which the attitude of a resident is more similar to that of her co-ethnic local neighbors than to that of her outgroup local neighbors (micro-level). Importantly, high local attitude alignment does not necessary imply a high difference between groups' mean attitudes at the district level (macro-level). Therefore, a second macro-level alignment measure we will use is the average local alignment score. Accordingly, we developed expectations and corresponding measures for *local alignment* (micro-level), *average local alignment* (macro-level) and *difference between mean attitude in groups* (macro-level).[5]

Let us consider local alignment first. In the RI-model, outgroup interactions are what tends to drive a wedge between ethnic groups: this is because outgroup interactions are more likely to increase disagreement with the ethnic outgroup. Therefore, we would expect that higher local outgroup exposure will facilitate the emergence of local alignment:

Expectation 2a) Agents who are locally more exposed to their outgroup exhibit higher scores of local alignment.

[5] Suppose for example a district where, in the East, residents from the majority ethnic group hold positive attitudes, whereas minority residents hold negative attitudes and that this pattern is reversed in the West. In the West, the majority has negative attitudes and the minority positive ones. Both in the East and the West we observe local alignment because residents tend to agree with their ingroup neighbors and disagree with their outgroup neighbors. However, in the district as a whole, attitudes are not necessarily more similar between same-group residents than between outgroup residents. We do not observe alignment at the district-level.

As we zoom out and focus on the district level, it is useful to remember that districts with higher average levels of outgroup exposure are not necessarily those with higher peaks in levels of local outgroup exposure (see "The spatial distribution of ethnic groups within Rotterdam". Fig. 2 and Table 1). It is therefore not trivial to derive which districts will develop higher average local alignment: those where on average outgroup exposure is higher, or those with more extreme scores of local outgroup exposure. Tentatively, we propose in expectation 2b that both district features are associated with higher local alignment:

Expectation 2b) Districts with higher levels of average outgroup exposure and districts with a higher share of agents highly exposed to their outgroup exhibit higher average local alignment.

Lastly, we examine the degree to which the attitude divide consistently runs between the two ethnic groups. Our intuition is that the average outgroup exposure (district-level) might have a non-monotonic effect on the attitude difference between groups. Sufficient exposure to the outgroup creates the conditions for repulsive influence, which sparks the extremization of attitudes. Sufficient exposure to the ingroup allows the groups to develop internal consensus. Both outgroup and ingroup exposure are thus needed to produce attitude differences between groups and consensus within. Thus, we can hypothesize an inverted U-shaped relation between average outgroup exposure and the attitude difference between groups (i.e., district-level alignment).

Expectation 2c) there is an inverted U-shaped effect of average outgroup exposure on the difference between mean attitude in groups.

It should be noted that we do not know which level of average outgroup exposure corresponds to the tipping point in expectation 2c. The tipping point might vary from district to district, as the geography of a district might influence where the tipping point is. Furthermore, the reasoning behind expectation 2c ignores the complex ways in which the relative size of the two groups may moderate the effect. For example, in districts where groups are of unequal size, the ethnic minority might be highly exposed to the majority, but not necessarily the other way around. In such districts, the relative size of the ethnic groups may thus moderate the effect of outgroup exposure in ways which are difficult to anticipate.

Ingredients of the ABM: agents' attributes, selection of interaction partner, social influence

Our implementation of the RI-model largely reflects previous implementations from the literature, with one main difference concerning how space and distances are modeled. Previous spatially-explicit implementations of the RI-model [9] assume a grid topology where agents–the grid cells–interact with one of the adjacent agents, chosen at random. In previous work it was therefore implied that the population density is constant across the map (because the grid is regular, by definition); and it implied that the probability of interaction is a step function of distance (positive for adjacent grid neighbors–and the same for all neighbors; and null for non-adjacent

Table 2 ABM pseudo-code

ABM scheduling

```
Initialize agents
For 200 time steps {
   For all agents i, taken in random sequence {
      Select an interaction partner j
      Calculate similarity weight between i and j
      Update the attitude of i and j
   }
   If system has converged, then stop
}
Compute outcome measures
```

neighbors). Our implementation relaxes these assumptions. First, instead of a grid, we assume that agents are placed on a continuous surface–this allows us to simulate realistic districts where population density varies between district locations. Second, we assume that the probability of interaction between agents is function of their distance in meters (instead of their grid-adjacency)–this allows us to better control the effect of distances on the dynamics of the RI-model.

This Section outlines our variant of the RI-model, introducing its entities, variables and scheduling. Each simulation run of the RI-model simulates the interactions and resulting attitude changes within a district. Simulations comprise two phases, initialization and attitude dynamics. Table 2 gives an overview of the model scheduling with pseudo-code, and each step will be explained in the next (Initialization and empirical seeding" and "Opinion dynamics). Simulation scripts are written for R 4.2.0 [44] and are available along with their documentation in a public GitHub repository.[6]

Initialization and empirical seeding

The initialization creates a population of agents as big as the population size of the modeled district of Rotterdam (see Table 1). Agents have three main features: their fixed geographic position, their fixed demographic attribute (i.e., their ethnic group membership, western or non-western), and their dynamic ethnic attitude. The geographic position and group membership of agents are based on empirical data; and the process of matching these features to empirical data is called "seeding" (or, more broadly, "empirical calibration" [14]).

To seed the model empirically we start with the data on the 100 by 100 m square areal units provided by Statistics Netherlands. Statistics Netherlands does not publicly provide statistics for square units with fewer than 5 residents–we thus exclude such areas from the tally reported in Table 1 and from our simulation experiment. For every remaining square unit, we create as many agents as there are residents. Agents' location is assumed to be the centroid of their square unit.[7] Agents'

[6] https://github.com/thomasfeliciani/scripts-NI-calibration.

[7] The proximity between two agents belonging to the same square unit level is assumed to be 52.14 m, which is the approximate average distance between all points in a 100 × 100 m square.

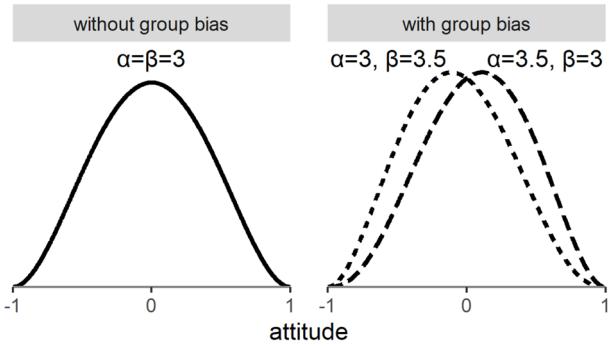

Fig. 3 Probability density functions for the initial distribution of ethnic attitudes in the simulated population. α and β are the parameters of the beta distribution, which we project to range between − 1 and + 1

ethnic group (non-western migration background or otherwise) is inferred from the observed proportion of residents with a non-western migration background in their square unit. The ethnic group membership of an agent i is coded $g_i = 1$ if i has a non-western migration background; $g_i = -1$ otherwise.

The last characteristic of agents that needs to be initialized is their ethnic attitude. We are interested in studying the conditions that facilitate the emergence of strong, polarized attitudes. We therefore assume that, at the start of the simulations, attitudes tend not to be extreme or strongly polarized. We achieve this by sampling initial attitudes from a beta distribution (with parameters α and β set to a values ≥ 1), and then transformed to range in $[-1,1]$. Figure 3 shows the parameterizations of α and β explored in this study. A case of particular interest is one where the two ethnic groups start out with a mild attitude bias: the initial attitude starts out on average slightly more positive in one group, and slightly more negative in the other (shown in Fig. 3, right panel). Our study will focus on this condition because it is plausible that ethnic groups might be biased regarding an ethnically-salient issue.

We want to check the robustness of our results to variation in the assumption that the two groups start out with an attitude bias–specifically, we want to see whether group bias is a necessary condition for alignment to emerge. Therefore, we replicate our simulations under the condition that there is no initial attitude difference between the two groups ($\alpha = \beta = 3$, left panel).[8]

Opinion dynamics

Within each of the 200 simulated time steps, all agents are selected in random sequence for an interaction event. An interaction consists in a calling agent i selecting an interaction partner j from the same district, and results in both i and j

[8] Simulations without initial bias mirror corresponding conditions for the RI-model in Flache [8], Mäs et al. [62].

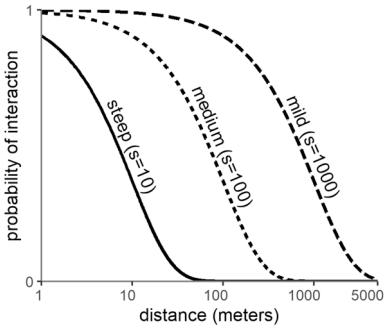

Fig. 4 Distance-decay functions

updating their ethnic attitude.[9] Thus, at every time step, $2 \times N$ attitude updates occur, all agents update their attitude at least once.

Selection of an interaction partner

We assume that intra-district interactions become less likely when two residents live further apart: there is a distance-decay in the intensity of interactions. Consistently with how we defined and measured proximity in the calculation of the index of outgroup exposure, here we assume that, for agent i, the probability to interact with agent j is function of the (normalized) distance to j as in Eq. 1 (which we show here again for clarity):

We measure d_{ij} in meters and assume s constant throughout the simulation run, but we run different simulations with different values of s, specifically $s \in \{10,100,1000\}$. These values are chosen for the magnitude of the geographical distances over which they allow interactions. As shown in Fig. 4, with $s = 10$, the probability of interaction between two agents drops very fast over increasing distances. Considering that the lowest level of aggregation is the 100 by 100 m square units in which each district is divided, with $s = 10$ the probability of interaction between individuals from different square units is virtually zero. This models strongly local interactions, where agents are hardly directly affected by the attitudes of others but their immediate neighbors. At the other extreme, $s = 1000$ allows interactions even between agents residing in distant district neighborhoods which, considering the size of these district, can be several kilometers apart. Parameter s, in sum, models the degree to which social influence can spillover from an agent's square unit to agents located farther out.

Varying s in our simulations serves two purposes. For one, we do not know what could be a realistic value for s in this context–that is, we do not know how the probability of having a potentially attitude-changing interaction decays over the distance

[9] In previous implementations of the RI-model, it was assumed that an interaction between a calling agent i and the interaction partner j results in the update of the attitude of i only. Our implementation differs slightly in that we assume both i and j update their attitude. This is done for the sake of computational efficiency, which was an important concern in our large-scale simulations.

from a potential interaction partner. Therefore, in our simulations we explore values across different orders of magnitude ($s \in \{10, 100, 1000\}$).

Second, parameter s can help us assess the implications of one of our modeling assumptions. We treat districts as if they were independent from one another: in our simulations we assume that interactions occur exclusively within districts and agents never interact with others outside of the district. In other words, we assume no spillover interactions across district boundaries. Like others have noted [5, 6], spillover interactions might affect the simulation results.[10] In a way, parameter s manipulates spillover interactions in our model—between localities of the same district rather than between localities from different districts: with higher s interactions occur more easily across larger distances, which signifies more spillover interactions between two localities. This allows us to vary s to gauge the effects of spillover interactions within the districts–which is a proxy for what we would expect from spillover interactions across district boundaries. Results of this examination are reported and discussed in Appendix A.

Attitude update: the repulsive influence model

Once i has selected j as interaction partner, we compute a weight w_{ij}: this weight determines whether the interaction is to lead to assimilation or repulsion and implements the effect of the similarity between the two interacting agents. The weight w_{ij} has range $[-1, +1]$ and captures the degree to which i and j are similar (or dissimilar) both in terms of attitude, $o \in [-1, +1]$, and of the ethnic group, $g \in \{-1, +1\}$. Similarly to previous implementations of this model [8, 38, 40, 45], the weight w_{ij} defines the sign and intensity of the social influence that j exerts on i. In particular, values of w_{ij} closer to the extremes (-1 and $+1$) result in stronger attitude shifts for i, whereas with $w_{ij} = 0$ no changes occur. Furthermore, for $0 < w_{ij} \leq 1$, the interaction triggers *assimilative* social influence, resulting in i updating her attitude to approach that of j. Conversely, $-1 \leq w_{ij} < 0$, triggers *repulsive* influence, meaning that the attitude of i further diverges from that of j. Formally:

$$w_{ij,t} = 1 - \left(\mathrm{H} \cdot \left| o_{j,t} - o_{i,t} \right| + (1 - \mathrm{H}) \cdot \left| g_j - g_i \right| \right) \tag{4}$$

Following Eq. 4, the weight w_{ij} is determined by the attitude difference between i and j ($|o_j - o_i| \in [0,2]$), their group difference ($|g_j - g_i| \in \{0,2\}$). We assume that the more similar agent i and agent j are, the more i's attitude will move towards j's attitude, or the less it will be repulsed by it, modelling homophily as the tendency to be more influenced by others who are more similar. A parameter H (which we call "homophily parameter") determines the 'source' of homophily: attitude similarity ($H = 1$), ethnic group membership ($H = 0$), or a mix of both ($0 < H < 1$). Generally,

[10] In [5, 6], the concern that cross-boundary interactions might matter is addressed by including in the simulation a 'buffer' of locations adjacent to the simulated district. This solution drastically increases the size of the simulated population and thereby the memory requirements of the simulation model. Because the districts of Rotterdam are much larger than those in [5, 6], simulating a buffer is not a practical solution in our case.

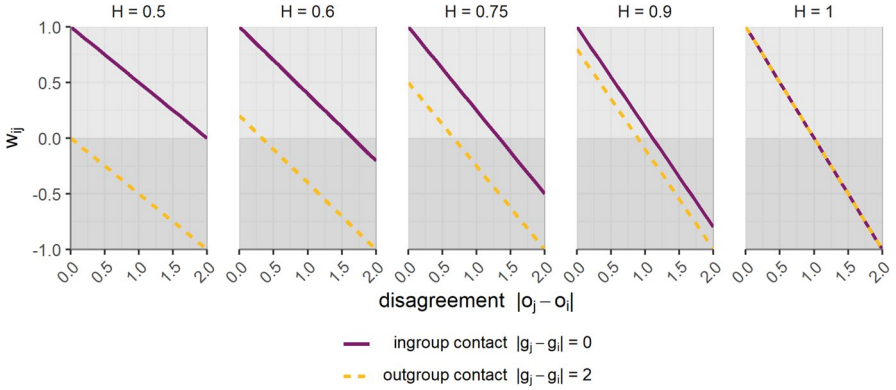

Fig. 5 An illustration of the effect of parameter H on the weight w_{ij} defined by Eq. 4. Five plots show, each for a different level of H, the relationship between disagreement (absolute attitude difference, on the X axis) and the weight w_{ij} (Y axis). Positive values of w_{ij} indicate assimilation (light-shaded areas of the plots); negative values indicate repulsion (dark-shaded areas). Two lines are shown: the solid line (dark, purple) refers to interactions between two agents of the same group (ingroup contact); and a dashed line (light, orange) refers to interactions between two agents of different groups (outgroup contact)

interactions between agents from the same group (i.e. when $|g_j - g_i| = 0$) require a lower degree of attitude disagreement for repulsion to happen; in other words, ingroup interactions have relatively more opportunities to result in attitude assimilation, whereas outgroup interactions (when $|g_j - g_i| = 2$) have relatively more chance to result in attitude repulsion. Following similar implementations in the literature [8, 9, 46], the homophily parameter H regulates what degree of attitude disagreement it takes for i and j to switch from assimilation ($w_{ij} > 0$) to repulsion ($w_{ij} < 0$), depending on whether i and j belong to the same ethnic group.[11] This is illustrated in Fig. 5.

When i and j belong to the same group (solid line), larger H means that smaller disagreement is necessary for repulsion to take place (observe in what regions of the X-axis the solid line falls in the darker region of the plot where $w_{ij} < 0$). It works the opposite way when i and j belong to different groups (dashed line): in this case, higher H means that disagreement needs to be larger for repulsion to take place. In sum, the larger the value of H, the lower the salience of ethnic group membership.

Figure 5 further helps us parameterize the homophily parameter H by showing that the values $H = 0.5$ and $H = 1$ are just outside of the theoretically relevant range. Setting H to 0.5 (left panel) signifies that when i and j belong to the same group only assimilation is possible, regardless of their attitude differences (the solid line always remains in the region $w_{ij} \geq 0$), and when they belong to different groups only repulsion is possible (dashed line fully in the region $w_{ij} \leq 0$). In other words, with $H = 0.5$ (or less) the direction of attitude changes is only determined by group membership, and not by attitude differences. At the other extreme, with $H = 1$ (right panel) all differences between the two groups disappear (solid and dashed lines are the same),

[11] This implementation of the homophily parameter H follows similar examples from the literature on the RI-model [8, 9, 46].

meaning that group membership does not play any role in determining the direction or magnitude of attitude changes–only attitude differences do.

Empirically we have no way of knowing what level of H more appropriately captures the dynamics of interactions among real people–although we can guess that the most appropriate value for H will vary among social settings and among individuals, and that H might also depend on the object of the attitude being considered, as well as the relevant group membership. For our purposes we can be content with exploring the system dynamics under values of H that are theoretically relevant, if not empirically grounded. Since our goal is to explore the consequences of ethnic segregation on attitude dynamics, the region of interest for us is in-between these two extremes (i.e. $0.5 < H < 1$), where group membership moderates the attitude dynamics. Therefore, in our simulation experiments we arbitrarily set H to 0.6 and 0.9 (second and fourth panels of Fig. 5) to represent scenarios where differences in group membership have a relatively strong vs relatively weak moderating role, respectively.

Once the weight w_{ij} is determined, we use it to calculate the raw attitude change of the two agents, $\Delta o_{i,t}$ and $\Delta o_{j,t}$ at time point:

$$\Delta o_{i,t} = \frac{1}{2} \cdot \left(o_{j,t} - o_{i,t}\right) \cdot w_{ij,t} \tag{5}$$

$$\Delta o_{j,t} = \frac{1}{2} \cdot \left(o_{i,t} - o_{j,t}\right) \cdot w_{ij,t} \tag{6}$$

At the end of the interaction, the attitudes of i and j are updated as follows:

$$\mathrm{o}_{i,t+1} = o_{i,t} + \Delta o_{i,t} \tag{7}$$

$$\mathrm{o}_{j,t+1} = o_{j,t} + \Delta o_{j,t} \tag{8}$$

A truncating function ensures that the updated attitudes do not exceed the range $[-1, +1]$.

Outcome measures

Our expectations from "Expectations" lay out a list of dependent variables: these are the model outcomes to be measured. The first is the number of interactions agents needed to develop an extreme attitude (from expectation 1a). We define "extreme" any attitude reaching value 1 or -1.[12] Depending on the sequence of interactions, an agent's attitude might repeatedly reach extremism, then be pulled back to more moderate levels, then extremize again, and so forth. We only consider agents' *time of first extremization*.

[12] To mitigate floating-point errors, we test for extremism as near-equivalence to these attitude extremes using a tolerance interval of $\pm 1.490116 \times 10^{-8}$.

Our second expectation (1b) concerns the emergence of *attitude polarization.* We measure this concept by calculating the standard deviation of the attitudes of agents.[13]

If we observe polarization, we are also interested in whether there is alignment of ethnic groups and ethnic attitudes both locally (expectation 2a), and in the district as a whole (expectations 2b and 2c). Starting with local alignment: to assess the extent to which residents from a specific ethnic group $\{-1,1\}$ are exposed to a specific attitude extreme [-1, 1] we propose a bivariate extension to Anselin's local indicator of spatial association, Moran's I (or simply "LISA"–see [47]). The bivariate LISA for agent i is defined as:

$$biv.LISA_i = \frac{1}{N} \cdot \frac{g_i - \bar{g}}{\sum_j \left(g_j - \bar{g}\right)^2} \cdot \sum_j p_{ij}\left(o_j - \bar{o}\right) \tag{9}$$

where $\bar{g}$ is the relative size of the minority in the district, and $\bar{o}$ is the average attitude.[14] Note that we are not interested in the sign of the correlation between ethnic group and attitude, because both positive and negative correlation imply alignment. Thus, we measure local alignment as taking the absolute score:

$$local\ alignment_i = \left|biv.LISA_i\right| \tag{10}$$

To measure alignment at the macro-level (expectations 2b, c), we first take the average of local alignment scores in a district to obtain *average local alignment*. A strong average local alignment score is a necessary but not sufficient condition for the attitude divide to run everywhere in the same way between the two ethnic groups (see expectation 2c). Thus, we also calculated the absolute difference between the mean attitude of the two ethnic groups. We refer to this measure as *difference between mean attitude in groups*. Since attitudes range in $[-1,+1]$, the mean difference ranges in [0,2].

[13] The literature offers more sophisticated measures of polarization designed to specifically capture the degree to which the attitude distribution becomes bimodal. One such measure is the polarization index used in Flache & Mäs [63] and follow-up work. However, this turned out to be computationally highly demanding for the large population sizes we explore here. Computation of the measure for small random samples of the agent population was possible. Based on this, we found that the standard deviation of attitudes very closely tracks the polarization index in all of our experiments. For better reliability, we thus decided to only report results for the standard deviation computed for the whole agent population rather than for random samples.

[14] In this literature, the weight matrix is commonly denoted w_{ij} – and not p_{ij}. However, we reserved w_{ij} for the influence weights in the RI-model (see Eq. 4). In Eq. 9 we thus use p_{ij} instead, since the weight p_{ij} is defined on the proximity matrix (see Eq. 1). Furthermore, N is formally defined as the row sum of the weight matrix. Because we row-standardized the proximity matrix it equals to the number of agents, N.

Table 3 Parameter space explored

Parameter	Values	Description
H	0.6 0.9	From Eq. 4, H determines the relative weight of differences in attitude vs in ethnic group in calculating the similarity between two interacting agents. For higher H, attitude differences count more than group differences (see Fig. 5)
Initial opinion distribution	$a=\beta=3$ (no initial group bias) for group 1: $a=3$, $\beta=3.5$; for group − 1 $a=3.5$, $\beta=3$ (initial group bias)	Defines the features of the initial attitude distribution. In the baseline we assume that the two groups start out with some initial attitude bias by imposing that the two groups originally have slightly different attitude means (see Fig. 3)
Distance decay function (s)	10 (steep) 100 1000 (mild)	A steep distance decay function means that agents interact only with other agents from their immediate surroundings, and hardly with any agent from farther than few hundred meters. A mild setting allows for good chances of interaction even between far-away agents (see Fig. 4)

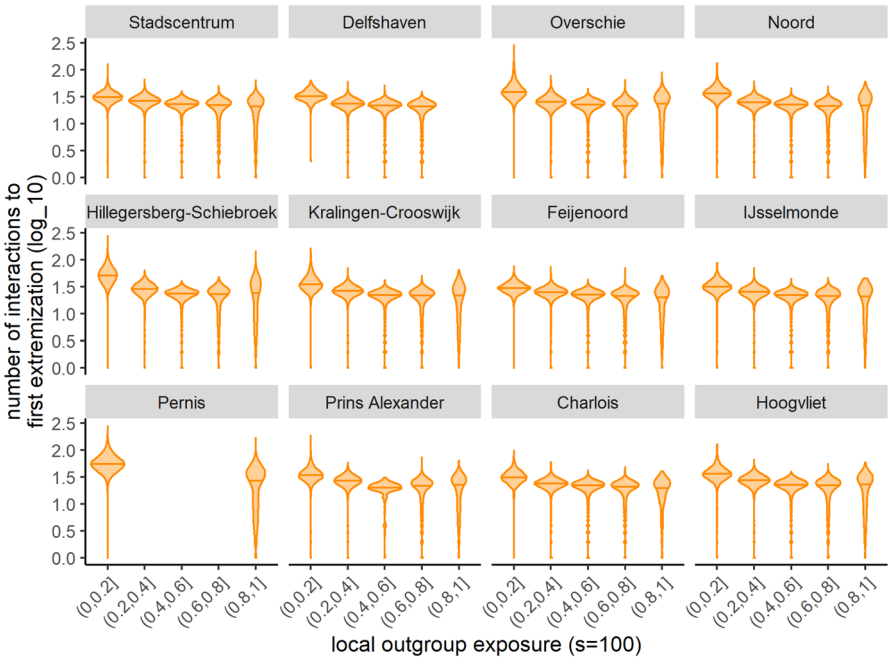

Fig. 6 For each district, we show the relationship between the local outgroup exposure ($s = 100$) of 500 random agents per simulation run and the number of interactions they had before eventually adopting an extreme attitude for the first time (log10 transformed). Baseline parameter configuration

Results

The parameter space we explored is summarized in Table 3. Results are averaged across 100 independent simulation runs per parameter configuration, for each of the 12 modeled districts.[15]

We ran 100 independent simulation runs per condition and for each of the 12 districts of Rotterdam. The underscored values show the 'baseline' configuration.

We denote one of the possible parameter combinations our *baseline configuration*–outlined in Table 2. To begin with, the baseline setting for the homophily parameter is $H = 0.6$. On the one hand, this setting is a conservative choice, as it implies that interactions between agents from the same group are unlikely to trigger mutual repulsion. One the other hand, the setting $H = 0.6$ reflects into the RI-model our theoretical assumption that ethnic membership moderates social influence. Predictions of the RI-model, turns out, are sensitive to the selected level of H–that is, to the strength of this assumed moderating role of ethnic membership on social

[15] To plot agent-level outcome measures (i.e. time of first extremization and local alignment) we randomly selected 500 agents per district, the same from each simulation. We resorted to a stratified sampling due to the memory requirements of the copiously large simulated populations. Figure captions indicate which plots are based on stratified samples.

influence: we therefore also present results assuming a much weaker moderating role ($H = 0.9$ see "Other parameter configurations" and Appendix A).

In the baseline configuration we further assume a middle value for s (the slope of the distance decay function), and that the two groups start out with slightly different average attitudes (initial group bias). We relax these assumptions in "Other parameter configurations" (and more fully in Appendix A) by exploring the predictions of the RI-model under parameter configurations alternative to the baseline.

In this baseline, all simulation runs eventually developed almost maximal attitude difference between groups: western agents reaching almost perfect consensus over one attitude extreme, and non-western agents gravitating towards the opposite attitude extreme. In the following "Outgroup exposure and time of first extremization (agent level)", "Outgroup exposure and alignment at the district level" we sequentially review our expectations in the light of simulation results in the baseline configuration.

Outgroup exposure and time of first extremization (agent level)

In the baseline parameter configuration, 99.98% of agents developed an extreme attitude ($o \sim \pm 1$) by the end of the simulation run. Figure 6 plots the number of interactions in which agents engaged before adopting an extreme attitude for the first time. Agents are grouped by district (panels) and by level of local outgroup exposure (levels on the X axis); the horizontal bar across the orange violin refers to the median time to first extremization in each violin. The wider the violin, the more agents reached extremization at that point in time.

Our first expectation (1a) is that agents who are more exposed to the outgroup become extremist faster, i.e., after fewer interactions. Figure 6 shows a more nuanced picture. On the one hand, it appears true for all districts that the agents least exposed to the outgroup extremize the last. On the other hand, those who extremize the earliest are generally not those with maximal local outgroup exposure. In fact, fastest extremization is typically observed for agents with an intermediate level of local outgroup exposure.

Our understanding of this result builds on the notion that attitude repulsion is strongest when interacting agents belong to different groups *and have very different attitudes*. Suppose there is a district where only one resident belongs to the minority group. In such a district, outgroup exposure is very low for majority members (corresponding to the left-most violin in our plot). For the minority agent, exposure is very high (right-most violin). Interactions involving the minority agent are most likely to result in repulsion. If the minority agent and her majority interaction partner have very different opinions, they both move to more extreme attitudes. However, majority members are surrounded by other majority agents, who, through assimilative influence, moderate the attitude of whomever has interacted with the minority agent. In other words, majority agents 'absorb' and thus slow down the extremizing force of repulsion. At the same time, for the minority individual it will be difficult to find an interaction partner with a strong and different attitude. Therefore, agents–majority and minority alike–located in parts of the district with very uneven group sizes

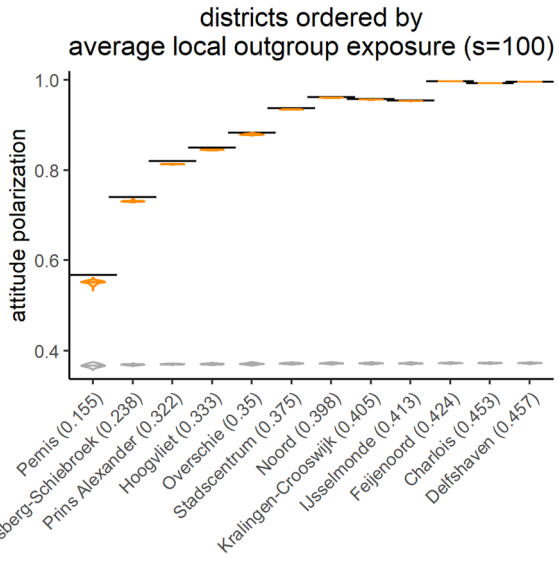

Fig. 7 Orange violins show attitude polarization at the end of the simulation runs. Attitude polarization is measured as the standard deviation of agents' attitude. The plot also shows attitude polarization at the start of the simulation (gray violins) and the maximum level of polarization (black line). Baseline parameter configuration

(and thus very high or very low levels of outgroup exposure) may take longer to become extremists. This is in line with the idea that agents' simultaneous exposure to both the ingroup and the outgroup facilitates the emergence of local alignment.

Outgroup exposure and polarization (district level)

At the agent level, extreme attitudes result from repulsion, which pushes the attitude of two interacting agents towards opposite extremes. At the district level, our expectation is that polarization is stronger in districts with higher average outgroup exposure (expectation 1b).

Violins in Fig. 7 plot the attitude polarization measured in the twelve districts ordered by average outgroup exposure: from Pernis, on the left, with the lowest average outgroup exposure, to Delfshaven, on the right, with the highest. We also plot the initial level of polarization (light gray): in the baseline configuration, the initial attitude bias causes the initial level of polarization to sit at about 0.39. Lastly, the plot also shows the level of polarization that would be observed under perfect district-level alignment, where all agents from one group are on one attitude extreme, and all agents of the other group are on the other extreme (black reference lines).

All runs in the baseline configuration converged to almost perfect district-level alignment (i.e., maximum attitude difference between groups). Correspondingly, Fig. 7 shows that attitude polarization (almost) reaches the maximum possible under

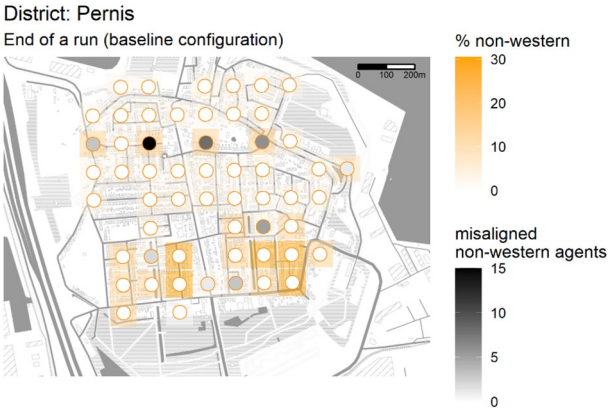

Fig. 8 Location of the misaligned simulated minority agents at the end of the simulation run ($t = 200$) in the district Pernis. Credit: basemap imagery from Stamen 2023

maximal district-level alignment (the orange violins are very short and very close to the black reference lines). However, districts differ in how much polarization there can be in case of perfect district-level alignment. This is shown by the black reference lines and can be explained by how we measure polarization and by the ethnic composition of each district. Polarization is operationalized as the standard deviation of agents' attitudes and is highest when the population is equally split into two opinion camps. Under perfect district-level alignment, this happens when the two ethnic groups are equally numerous, like is approximately the case for Charlois and Delfshaven. Conversely, unbalanced ethnic compositions result in uneven opinion camps and thus lower potential for polarization: this is the case for districts such as Pernis, characterized by a larger western majority and a much smaller non-western minority. Crucially, relative group sizes are also related to a district's average outgroup exposure (see also Table 1): and this is why we find districts like Pernis on the left side of Fig. 7 (uneven ethnic group sizes; low average outgroup exposure; low expected polarization in case of perfect alignment); and districts like Delfshaven on the right (even ethnic group sizes; high average outgroup exposure; high expected polarization under perfect alignment).

Consistently with expectation 1b, polarization is lower than the black reference line in districts with low average outgroup exposure (left of Fig. 7). In other words, simulations of districts like Pernis generate less polarization than would be observed in case of perfect district-level alignment. The reason for this discrepancy appears to be presence of a few "misaligned" ethnic minority agents: "misaligned" in the sense that they sided with outgroup members in their district by adopting their attitude extreme. Misaligned minority members make the opinion camps even more unbalanced, thereby reducing the level of polarization.

We think that two conditions facilitate the misalignment of an agent like we see happening in Pernis: (1) an initial attitude that, by chance, is closer to the average attitude of the outgroup than it is to the ingroup's; (2) strong exposure to the outgroup. This second condition is more likely met for minority agents, as they are the

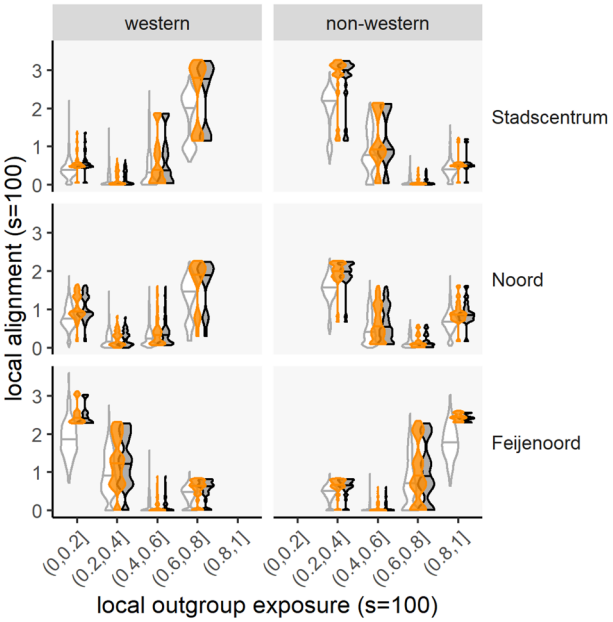

Fig. 9 The effect of local outgroup exposure ($s = 100$) on local alignment scores ($s = 100$) for a stratified sample of 500 random agents per simulation run. Alignment scores are measured at the start of the simulation runs (gray violins) and at the end (orange violins). As a reference, black violins represent local alignment scores in case of perfect macro-level alignment. Data are disaggregated by district and ethnic group. Baseline parameter configuration

ones more likely isolated in ethnic-majority parts of the district: this is the reason why we find misaligned agents among the minority ethnic group and not in the majority group. Figure 8 shows a map of Pernis with the square units where minority residents are more likely to side with the majority. Misaligned ethnic minority residents emerge in areas where local outgroup exposure is high.

Local outgroup exposure and local alignment

For the next group of expectations, we focus on the degree to which attitude polarization occurs between groups – a phenomenon we called district-level alignment. With expectation 2a we have conjectured that agents' local exposure to outgroup members fosters local alignment. In Fig. 9 results are disaggregated by ethnic membership and only include three representative districts; the full table with all districts is found in Appendix B.

We start reading these plots from the distribution of alignment scores at the start of the simulation (light gray violins). First of all, these violins appear to be approximately mirrored between the two ethnic groups. This is because residents who live in the same location but who belong to a different ethnic group have opposite values of relative outgroup exposure, while sharing the same degree of local alignment.

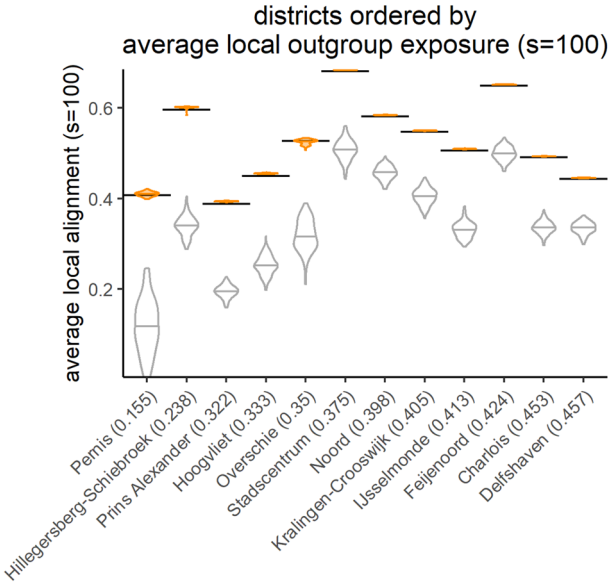

Fig. 10 Average local alignment (s = 100) at the end of the simulation run (orange violins). The plot also shows average local alignment at start of the simulation (gray violins), and under the condition of maximum mean attitude difference between groups (black line). Baseline parameter configuration

Secondly, alignment scores follow a U-shaped distribution. This U-shape results from the operationalization of local alignment. It follows from Eq. 9 ("Outcome measures") that in the situation where disagreement between groups is maximal–as in our simulations–the local alignment score tends to zero for agents whose outgroup exposure approaches the proportion of outgroup members in the district.

Now, our expectation 2a is of a positive relationship between local outgroup exposure and local alignment. However, Fig. 9 shows that this is not the case: at the end of the simulations (orange violins), alignment has increased by a proportionate amount both where local outgroup exposure is low and where it is high. Not only that, because our runs have all (almost) converged to perfect district-level alignment, local alignment scores have reached their theoretical maximum, represented in Fig. 9 by the black reference violins.

Outgroup exposure and alignment at the district level

Next, we want to see if there are differences between districts in the *average local alignment* (expectation 2b). We initially expected expectation 2a to scale to the district level and that we find that average local alignment is stronger in districts where the outgroup exposure is on average higher (and where there are more highly exposed residents). This is however now even more in question,

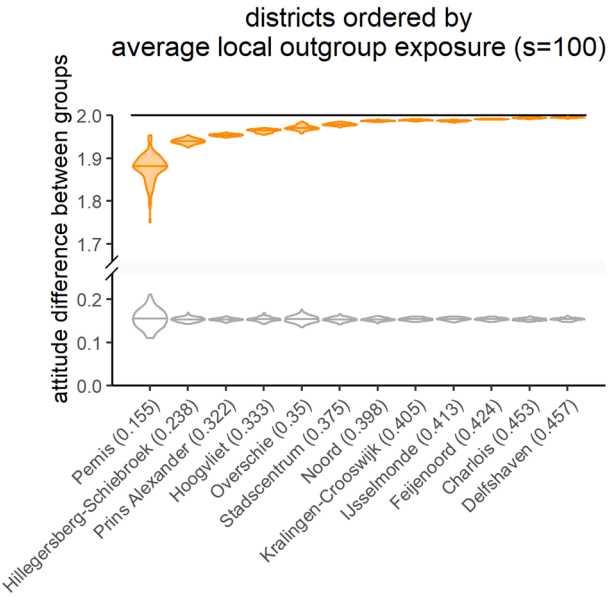

Fig. 11 Difference in mean attitudes between groups in the simulated districts of Rotterdam. Measurements are taken at the start of the simulation (light gray) and at the end ($t = 200$; orange). A black line marks the maximum theoretical district-level alignment (=2). Baseline parameter configuration

considering that simulation runs in the baseline have converged to almost perfect district-level alignment and since we have not found support for expectation 2a.

Violins in Fig. 10 show the distribution of average local alignment generated by the RI-model. Gray violins show that districts vary in average local alignment at the outset of the simulation run, and that initial local alignment tends to be lower in districts with low average outgroup exposure (such as Pernis, Prins Alexander and Hoogvliet). Orange violins show that this difference between districts tends to disappear by the end of the simulation run: at $t = 200$, average local alignment has increased in all districts, up to about their maximum (black line).

We find no discernible relationship at the district level between outgroup exposure and average local alignment–neither in Fig. 10, where districts are ordered by average outgroup exposure, nor by ordering districts by the proportion of agents who are highly exposed to the outgroup.

Last we examine results for expectation 2c. We expected an inverted *U* shaped relationship between average outgroup exposure and the difference in means between groups–but simulations prove us wrong. Figure 11 shows instead a monotonically positive trend.

As we have already discussed, all simulation runs converged to almost perfect district-level alignment. Simulation runs that fell short of reaching perfect alignment are of districts with low average outgroup exposure (left side of Fig. 11). In "Outgroup exposure and polarization (district level)" we have attributed this discrepancy to the presence of a few misaligned minority members in these districts.

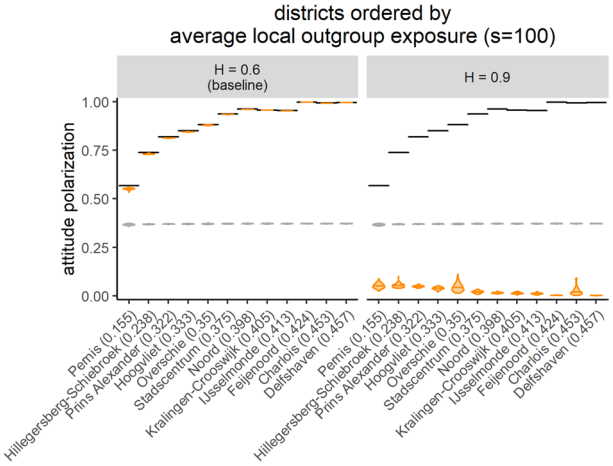

Fig. 12 Attitude polarization in the simulated districts of Rotterdam at $t=0$ (light gray) and $t=200$ (orange). All parameters except for H are set to their baseline value

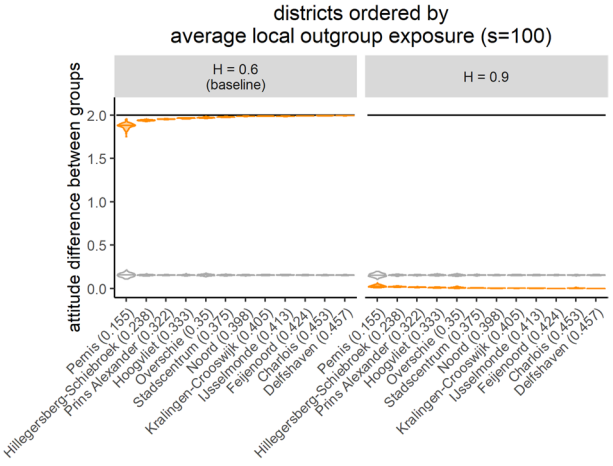

Fig. 13 District-level alignment (measured as attitude difference between groups) in the simulated districts of Rotterdam at $t=0$ (light gray) and $t=200$ (orange). All parameters except for H are set to their baseline value

Other parameter configurations

We have shown that RI-model simulations from the baseline configuration are generally characterized by strong polarization and strong attitude difference between groups. Alternative parameterizations of the RI-model can produce different outcomes.

One important difference can be observed when we vary the homophily parameter H. In the baseline configuration we have assumed a strong moderating role

of ethnic group membership, which we have captured by setting $H = 0.6$. An alternative setting is $H = 0.9$, where this moderating role is much weaker. With $H = 0.6$ we observed that RI-model simulations tend to generate attitude polarization between groups. This does not happen at all with $H = 0.9$. For instance, Fig. 12 shows that attitude polarization at $t = 0$ (gray) increases over time with $H = 0.6$ (orange–$t = 200$) and it decreases to the point of almost vanishing with $H = 0.9$. Likewise, Fig. 13 shows that attitude differences between groups–wide with $H = 0.6$–are nihil when we set $H = 0.9$. We further observed that the few agents who developed an extreme attitude at all with $H = 0.9$ did so in their very first interactions, where the attitude bias at the start of the simulation allowed for occasional encounters between agents who, by chance, had started out with sufficiently large attitude differences. Virtually all agents became more moderate after the first few interactions and eventually arrived at a global consensus on a moderate attitude.

Appendix A investigates this effect of H further, exploring where between $H = 0.6$ and $H = 0.9$ the RI-model starts generating attitude consensus rather than polarization between groups. Here it suffices to remark that the results from Figs. 12 and 13 show the importance of the assumption that ethnic membership moderates the interpersonal influence dynamics.

Another important assumption for the predictions of the RI-model concerns the attitude difference between groups at the start of the simulation. In our baseline configurations we assumed that there is some degree of attitude difference between the two ethnic groups right from the start of the simulation–an initial 'group bias'. When there is no initial group bias the RI-model may generate attitude distributions other than polarization and strong district-level alignment: it may also generate moderate consensus and polarization within groups (where the attitude divide cuts across the ethnic boundary–see Appendix A, Figs. 15 and 16). Which equilibrium is more prevalent depends on which district is being simulated. In the investigated districts characterized by relatively low outgroup exposure (i.e. high segregation), attitudes tend to become polarized (Fig. 15), but not necessarily aligned with group membership (Fig. 16). For investigated districts with moderate levels of average outgroup exposure, high attitude difference between groups remains the most likely end state and hence polarization does not reach its maximum. For districts with relatively high levels of average outgroup exposure, we observed an even more complex pattern. Attitude polarization could either start to cut across ethnic group membership and become nearly maximal or residents reach could consensus over ethnic attitudes, or a split of attitudes along ethnic boundaries could occur. This confirms what we have already observed: that the composition and geography of the district can affect, ceteris paribus, the outcomes of the simulation model. This also puts into perspective results obtained by earlier work with the RI-model using simper spatial structures [8, 9]. A prevalent finding in earlier work was that more segregation can reduce polarization; here, using more realistic spatial structures and segregation patterns we paint a more nuanced picture.

Conclusion

We examined the link between patterns in ethnic-residential segregation and the spatial distribution of ethnic attitudes. We have demonstrated that this link does not necessarily hinge on neighborhood effects. Rather, it can in principle emerge from a small set of assumptions: first, that ethnicity moderates the dynamics of interpersonal attitude influence – that is, individuals can influence each other's attitudes differently, depending on whether they share the same ethnic membership. Second, that interactions are local: individuals are more likely to interact and influence the attitude of people who live close to them, and less likely to influence people farther away. Third, that the spatial distribution of ethnic groups is uneven (i.e. groups are spatially segregated).

We explored this idea building these assumptions into an agent-based model of interpersonal social influence, the RI-model. In our simulation experiment we reproduced the spatial arrangement of two ethnic groups as observed in 12 districts of Rotterdam (natives and western on the one hand, and non-western on the other). These districts vary in segregation patterns and relative size of the two ethnic groups, and as such are an ideal testing ground for observing how different ethnic compositions affect the emerging attitudes under the dynamics of the RI-model.

We have focused on two aspects of the distribution of ethnic attitudes: the degree to which attitudes become polarized, and the degree to which the attitude divide overlaps with the ethnic divide, leading the two ethnic groups to hold opposite ethnic attitudes. In both cases, and at various levels of aggregation, we found that model dynamics led under a baseline configuration to almost maximal attitude differences between groups across all districts. We further found that the ethnic composition of a district has a big impact on the potential for attitude polarization in the population as a whole and for local alignment of attitudes, but not for the extent of attitude differences between groups. In general, our simulation experiment supports our intuition that micro-level influence processes can link ethnic-residential patterns to patterns in ethnic attitudes, warranting further empirical research on the subject.

Limitations and simplifying modeling assumptions

Like for all simulation research, the soundness and generalizability of these results are just as good as the assumptions of the model. This speaks to the limitations of our work which mostly derive from the simplifying, unrealistic assumptions of the RI-model. These simplifying assumptions reflect the goal of our study, which is not (yet) to make realistic predictions e.g. about how attitudes are distributed in a city district–this would indeed require relaxing a number of simplifying assumptions. Rather, our goal was to demonstrate the complexity of the relationship between ethnic residential segregation and attitude polarization–a goal that requires theoretical parsimony and justifies the simplifying assumptions.

One of the simplifying assumptions concerns who can influence an individual's attitudes. We wanted to learn how ethnic residential segregation affects the RI-model: therefore, in our model we only considered neighborly interactions, where

an individual's attitude can only be influence by her neighbors–and the farther the place of residence of the neighbors, the lower the likelihood that they will interact and have an influence. In the real world, we can of course meet with (and be influenced by) others regardless of residences' proximity, simply because we can travel; or communicate across distances (e.g. via social media).

Further simplifications follow from the micro-level processes of attitude change described by the RI-model. For instance, in case of encounters between neighbors with different ethnic membership, individuals are assumed to adjust their attitudes based on the attitudes on the same issue held by the neighbor. This neglects intergroup-contact as another important mechanism for how a specific class of attitudes, attitudes about the outgroup, like prejudices, can change in between-group interactions. The core mechanism of contact theory [48, 49] is the generalization from positively experienced interpersonal contact with an outgroup member towards positive attitudes about that outgroup as a whole. Intergroup contact and the RI-model describe similar but different micro-level processes and may thus have different implications for the effects of residential segregation on the distribution of attitudes about outgroups in a population. Specifically, the social influence mechanism of the RI-model can be considered more general, as it describes the simultaneous effects of interactions both within and between groups and it allows modelling changes with regard to any ethnically salient attitude, not just attitudes about outgroups. We leave it to future work to explore the interplay of social influence and intergroup contact in ethnically diverse spatial settings.

Relatedly, we only studied one model of social influence, the RI-model. Human interactions (and attitude dynamics specifically) are governed by a multitude of processes, often at play simultaneously. One assumption of the RI-model that is of great significance for the dynamics we observe is that of repulsive influence: agents change their attitudes so as to increase opinion disagreement with negatively evaluated sources of influence. It is an outstanding question in empirical research on social influence dynamics whether such so-called "boomerang effects" actually occur in real-life social influence. Under certain laboratory settings no evidence was found [50]. But more sophisticated online experiments [51] and several empirical studies of field settings [52–54] provide some support for a boomerang effect. Future work needs to explore whether alternative processes of social influence (e.g. [55, 56]) other than the ones in the RI-model, or a different mix of processes, may lead to different conclusions.

A further simplification is that we assumed that interactions can only occur within districts, and never across district boundaries. On the one hand, there are some advantages to this assumption: for instance, it allows us to treat districts as independent from each other–and to compare them, which gives us insight into the effects of spatial characteristics of districts on polarization of ethnic attitudes. Furthermore, the demanding computational requirements of large scale simulation models make it prohibitively expensive to simulate interactions at the scale of the whole city. On the other hand, the implausibility of this assumption is mitigated by the geography of Rotterdam, where districts are often physically separated; and we argue that it is of little consequence for the interpretation of the model results (see Appendix A).

Another limitation concerns the narrow definition of segregation adopted here. We have taken segregation to mean agents' relative exposure to their outgroup: the lower the exposure, the stronger the segregation. On the one hand, this definition allowed us to compare districts combining the two dimensions of segregation we deemed relevant: the degree of relative proximity to one's outgroup, and its relative size. On the other hand, our results point at some differences between districts which are not explained by outgroup exposure. This signals that we have not captured all the attributes of the district topology or of the spatial arrangement of the ethnic groups which are responsible for these differences.

Lastly, another limitation is the implementation of the distribution of initial attitudes in the population. In this work we have assumed that the distribution of attitudes at the start of the simulation is bell-shaped. We have then explored the consequences of assuming, or not assuming, that there is a slight difference between the average attitudes of the two groups at the start of the simulation. The presence of this initial group bias plays a decisive role for the simulation dynamics (see "Other parameter configurations" and a fuller explanation in Appendix A). This indicates clear directions for future work: first, to fine tune this insight and identify more precisely just how much initial bias is sufficient to produce observable changes in the degree of polarization and alignment in the districts we simulated. Second, whether and how our results also hinge on the assumption that the distribution of attitudes at the start of the simulation is unimodal and bell-shaped.

Conclusions and directions for future work

In our experiment even the most intuitive parameter manipulations generated results which were more nuanced and complex than we could anticipate prior to running the experiment. This teaches us two lessons: first, that we would be ill-advised if we based our expectations on the interactions between elements of a complex system solely on our intuition. Second, that increasing model realism and decreasing its abstraction can lead to unexpected insights and is an exercise often worth pursuing.

Our focus was on the spatial extension of the RI-model that allows us to study how ethnic residential segregation affects its dynamics: this is the area where we increased model realism and decreased its abstraction. First of all, compared to previous work that relied on abstract, stylized segregation patterns [8, 9], here we seeded the input of the RI-model model with empirical data on the spatial arrangement of ethnic groups. Second, we moved from the grid topology used in previous implementations to a continuous surface, which in turn allowed us to study districts of different shapes with internally varying population densities and to explore more realistic distance functions. Third, we examined the model outcomes more comprehensively, focusing both on how the dynamics vary among district locations, and how they evolve through time (as opposed to only examining outcomes at one time point). All of these innovations proved necessary to surface and understand previously unnoticed dynamics in the RI-model and to observe unknown interactions between the spatial arrangements of agents and groups and the well-studied parameters of the RI-model.

One of these previously unnoticed dynamics and interactions concerns a fundamental prediction of the RI-model, that ethnic segregation is negatively related to polarization. The intuition behind it is that more exposure to outgroups (less segregation) increases the potential for disagreement to develop between groups, which ultimately leads to attitude polarization and, *to varying degrees*, alignment of attitudes and ethnic membership [8, 9]. With the spatially-extended version of the RI-model we can now paint a more complete picture. By comparing our simulated district in our baseline parametrization (i.e. where groups start off with an initial bias in mean ethnic attitudes and ethnic group membership is important) we show that the predicted degree of polarization depends not only – or not so much – on the degree of segregation, but rather on the relative size of the two ethnic groups, and that stronger polarization emerges in districts where the two groups have similar sizes.

Looking at the model outcomes more comprehensively surfaced new insights in how the polarization dynamics unfold at the micro level, too. In research based on the RI-model, it was previously assumed that opportunity of contact with the outgroup is the spark needed for agents to adopt extreme attitude and for polarization to emerge. Here, by looking how polarization evolves through time and in different parts of the district, we show that agents polarize more and faster in parts of districts where ethnic groups are *mixed*: this means that agents polarize (faster) when they have plenty of opportunities to interact *with their ethnic ingroup* as well their outgroup.

As our additional robustness tests showed, when group membership does not play a major role in the social influence dynamics, district residents most often reached consensus over the ethnic attitude. Moreover, when groups do not differ in their mean ethnic attitude at the start of the simulation (but when group membership does play a role in the influence dynamics), we observed a lot more variation across districts, and within districts across simulation runs, in our outcome measures. Given our relatively low number of investigated districts and that we also observed deviations from the above described pattern, this warrants a follow-up study to examine in more detail how patterns of segregation drive model outcomes.

In conclusion, future work on the RI-model can include the relaxing of some of the simplifying modeling assumptions and a more fine-grained characterization of the segregation patterns in the modeled districts. We note however that both lines of research require expanding the simulation model by introducing new modeling parameters, introducing new independent variables and thus running larger simulation experiments. Here, the high computational requirements of large-scale simulation models constitute an important bottleneck.

Other authors have commented on the issue of the prohibitive computational requirements of large-scale ABMs, particularly for simulation experiments with high-dimensional parameter spaces (requiring very long CPU time), and particularly for spatially-explicit models (requiring the calculation and storage of very large distance matrices) [57–59]. New ABM research can be enabled by the study of generalizable strategies to optimize large-scale, spatially-explicit ABMs and ease the computational constraints. This points to methodological improvements as a promising–and, for us, necessary–direction for future work.

Appendix A

In this appendix we examine the dynamics of the model outside of the baseline parameter configuration. As a reminder, we have introduced three model parameters (see overview in Table 3): the homophily parameter (denoted H); the presence of ethnic bias in the attitude distribution at the start of the simulation; and the slope of the distance decay function (captured by s). Where we examine them in sequence, comparing ceteris paribus the baseline configuration with the alternative parameterizations.

Homophily parameter, H

The homophily parameter H determines the 'source' of homophily: attitude similarity ($H=1$), ethnic group membership ($H=0$), or a mix of both ($0<H<1$). Specifically, following Eq. 4 ("Opinion dynamics"), H moderates how attitude differences between two agents and their ethnic group membership affect the direction and magnitude of attitude changes. In our baseline configuration we have assumed that $H=0.6$, meaning that direction and magnitude of attitude changes are strongly influenced by the ethnic membership. With $H=0.6$, if two interacting agents i and j belong to the same ethnic group, they need to be in very wide disagreement for repulsion to take place; and conversely, if they belong to different groups very little disagreement is sufficient.

In results "Other parameter configurations" we have compared the model results from $H=0.6$ (baseline configuration) with $H=0.9$. While with $H=0.6$ the RI-model generates strong attitude polarization and large attitude differences between groups, with $H=0.9$ it generates instead attitude consensus within and between groups, and in all districts (see Figs. 12 and 13). This large difference between the two values of H raises curiosity about the model behavior when H is set to intermediate levels. In particular, we do not know whether there is a threshold level of H below which the model generates alignment and polarization and beyond which it generates moderate consensus; whether the transition from alignment and polarization to consensus is rather more gradual as H increases; or whether with intermediate levels of H we observe altogether different model behaviors. We ran some additional simulations to answer these questions. The new simulations focus on the districts Pernis and Overschie, which we chose for convenience (they are the smallest districts by population size) and because they show large differences in group composition and outgroup exposure. For both districts we explored levels of H from 0.6 to 0.9 in steps of 0.05, setting all other parameters to their baseline configuration, and running five simulation runs per condition. Results are shown in Fig. 14.

The effect of parameter H is monotonically negative in both districts, confirming that higher H (that is, when ethnic membership is relatively less salient) is more conducive to attitude consensus within and between groups. We also notice a difference between the two districts. In Overschie (left panels), there seems to be a threshold in the region $0.65<H<0.7$: below this threshold we find strong polarization and alignment; above it, consensus. By contrast, in Pernis

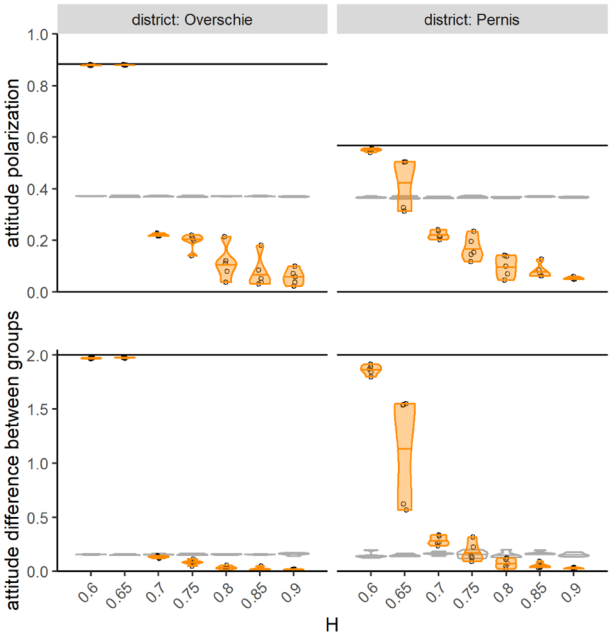

Fig. 14 Additional simulation runs to explore the effect of H between the values 0.6 and 0.9, in steps of 0.05. Results are based on five runs per condition, and we simulated two districts (Overschie, left; and Pernis, right). The panels show the degree of attitude polarization (top panels) and the attitude difference between groups (bottom panels). Inside each panel, orange violins and black circle markers show these measures at the end of the simulation runs ($t=200$); gray violins show them at the start of the simulation ($t=0$); and a black line marks their theoretical maxima in case of maximum polarization and perfect district-level alignment. All parameters except for H are set to their baseline configuration

(right panels), this transition is somewhat more gradual and happens at exactly $H=0.65$. What can be learned from this is that the RI-model predicts differences between districts in how the role of group membership H is mapped on the two model outcomes – consensus versus alignment and polarization. There are levels of H (here, $H=0.65$) where consensus can emerge in one district (here, Pernis) and not in another district (Overschie). These differences between districts can reasonably be attributed to either their population size, group composition, or degree of segregation.

In conclusion, these results tell us that in situations where ethnicity has a strong moderating role on attitude dynamics (e.g. attitudes towards migration policies), the RI-model predicts that attitudes will polarize and align with the ethnic divide. Conversely, domains where ethnic membership is sufficiently irrelevant will see ethnic groups converge towards a shared, moderate consensus.

Initial group bias

In our baseline configuration agents start out with some degree of alignment. We have observed how interactions in the RI-model amplify this initial group bias until the difference between the average attitude of the two groups almost reaches its maximum. It is interesting now to observe if attitudes polarize and the attitude gap between groups emerges when there is no initial alignment.

The left panel of Figs. 15 and 16 shows that, generally (though to varying degrees and with a few exceptions) both attitude polarization and attitude difference between groups are higher at the end of the simulation (orange violins) than they were at the start (gray).

On the one hand, this result confirms that initial group-level attitude bias is not a necessary condition for the emergence of attitude polarization. Quite the contrary: Fig. 15 shows that some districts can achieve even *stronger* polarization if there is no initial group bias. The reason for this is that polarization is maximal when the population is split equally into two opinion camps; and for districts where the ethnic groups have uneven sizes, polarization can only be maximal if it does not align with ethnic membership. Since initial group bias generates alignment, it also prevents maximal polarization in districts with uneven group sizes. This is why polarization can peak only without initial group bias in districts with very uneven group sizes (see e.g. Prins Alexander in Fig. 15).

On the other hand, the initial group bias facilitates the emergence of attitude differences between groups. This can be seen in the left panel of Fig. 16, where simulations often produced levels of between-groups attitude differences lower than in the baseline (right panel).

A combination of factors might explain why between-groups alignment is lower when there is no initial group bias. The first and most important factor is that, without initial group bias, different parts of a district might simultaneously converge

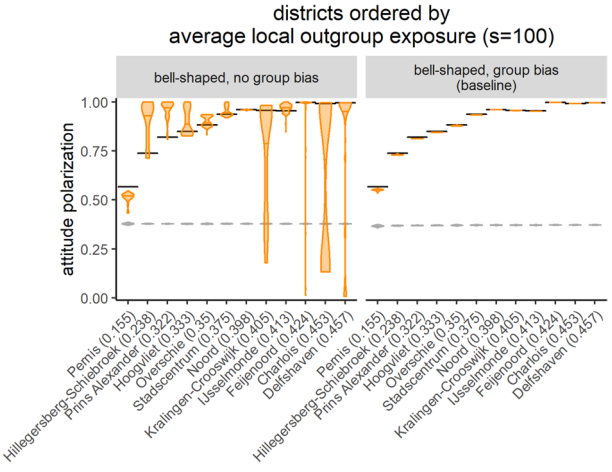

Fig. 15 Attitude polarization in the simulated districts of Rotterdam at $t=0$ (light gray) and $t=200$ (orange). All parameters except for the initial attitude distribution are set to their base-line value

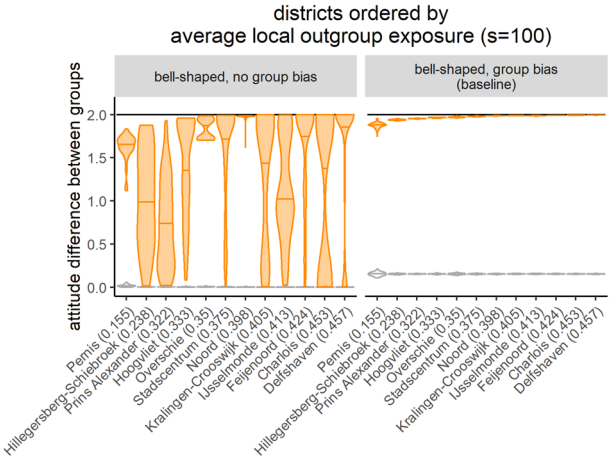

Fig. 16 District-level alignment (measured as attitude difference between groups) in the simulated districts of Rotterdam at $t=0$ (light gray) and $t=200$ (orange). All parameters except for the initial attitude distribution are set to their baseline value

towards opposite equilibria: in some parts of the district natives and western might converge towards attitude + 1, while non-western towards -1; and vice-versa in some other parts of the district. This situation describes those runs in Figs. 15–16 (left panel) with very high attitude polarization but little to nil between-groups attitude differences.

A second and related factor is that the dynamics are just faster if there is group bias at the start of the simulation, so much so that many of the runs without initial bias had not yet fully converged by the end of the simulation ($t=200$). Because different parts of the district might start converging towards opposite equilibria, more interactions are needed for agents to coordinate with their group at the district level. In Figs. 15 and 16 we are thus comparing runs where between-groups polarization has had a head start and is well underway (right panel), with runs where it has just begun to emerge, and might eventually be fully achieved (left panel).

A third and last factor is the presence of a few simulations runs that converged to attitude consensus rather than polarization—and the fact that here attitudes did not polarize at all explains why we do not find attitude differences between groups. We can see these 'consensus runs' in a few districts in Fig. 15 (left panel), characterized by very low levels of attitude polarization.

This leaves us with a further puzzle. There is no obvious reason why, without initial bias, only some districts produce simulation runs converging to consensus (see e.g. Charlois); only some to polarization without strong attitude difference between groups (e.g. Ijsselmonde); and others converge to strong polarization and strong average difference between groups (e.g. Noord). Overall, the lack of initial group bias makes it harder for agents to align their attitudes with their ethnic

membership–and this condition seems to make the spatial features of the district more relevant for determining which equilibria will emerge. What these features exactly are and why they affect the attitude dynamics in the way they do is a research question for future work. Tentatively, we propose that an explanation must lie in the interaction between the local ethnic composition of these districts, their population size and the relative size of their two ethnic groups.

Slope of the distance decay function, s

In essence, parameter s defines the extent to which social influence can spill over from an agent's location to other locations father away. Compared to the baseline ($s = 100$), with a milder slope ($s = 1000$) geographical proximity weighs far less on the probability of interaction. This increases the influence between agents who live some kilometers apart, e.g. in different neighborhoods from the same city district. By contrast, a steeper slope ($s = 10$) means that interactions are far less likely to occur between agents from more than one or two hundred meters apart – which effectively prevents social influence from spilling over even to nearby locations.

In our results, attitudes polarize and alignment emerges regardless of the slope, though to varying degrees, and with a few considerations to be made. With strictly local interactions ($s = 10$) we find more misaligned agents, especially in districts with low average outgroup exposure (such as Pernis). Unlike we saw in our baseline configuration, here both ethnic minority and majority agents become misaligned. The reason for misalignment is that strongly local interactions facilitate the emergence of opposite local alignment "spins" in different parts of the district. When interactions are more local, there are more agents who find themselves alone (or among few) and isolated in interaction neighborhoods where they

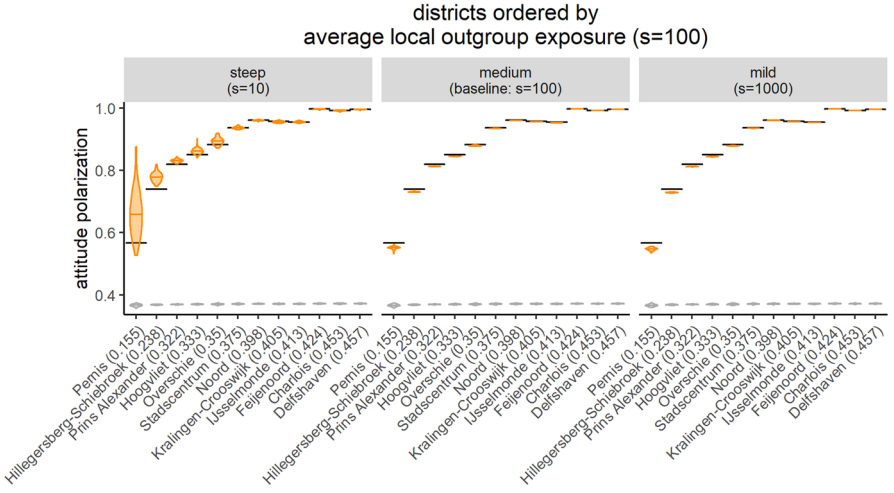

Fig. 17 Attitude polarization in the simulated districts of Rotterdam at $t = 0$ (light gray) and $t = 200$ (orange). All parameters except for s are set to their baseline value

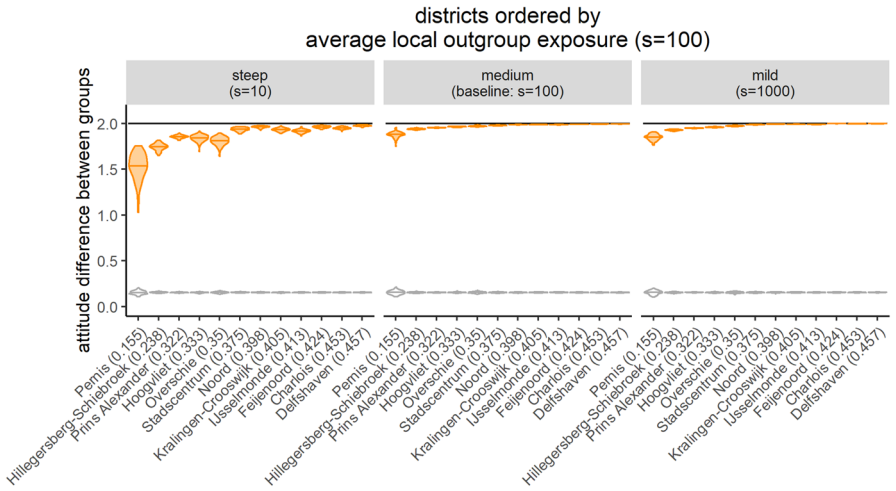

Fig. 18 District-level alignment (measured as attitude difference between groups) in the simulated districts of Rotterdam at $t=0$ (light gray) and $t=200$ (orange). All parameters except for s are set to their baseline value

are the ethnic minority. As we have explained in "Outgroup exposure and polarization (district level)", strong exposure to the outgroup is an important favorable condition for the emergence of misalignment. The consequence of higher misalignment can be seen in the figures below: with $s=10$, despite higher polarization rates (Fig. 17) and attitude difference between groups (Fig. 18) are lower than with $s=100$ and $s=1000$.

One last observation to be made is that, thanks to the more widespread misalignment, with $s=10$ attitude polarization reaches levels higher than the reference line (Fig. 17). This is further support for the idea of a trade-off between maximum polarization and maximum alignment: where misaligned agents create evenly-sized attitude factions—which increases polarization—they do so by siding with their ethnic outgroup—which decreases local alignment and the attitude gap between the ethnic groups. As we observed previously for simulations without initial group bias, also with strictly local interactions we find polarization that splits the population along other lines than group boundaries, allowing for a higher degree of overall attitude polarization.

Appendix B

Here we include an extended version of Fig. 9, discussed in "Local outgroup exposure and local alignment", and here including all simulated districts. Like Fig. 9, Fig. 19 shows—for all districts—that local alignment at the end of the simulation run (orange violins) is proportionally higher than at the start (gray).

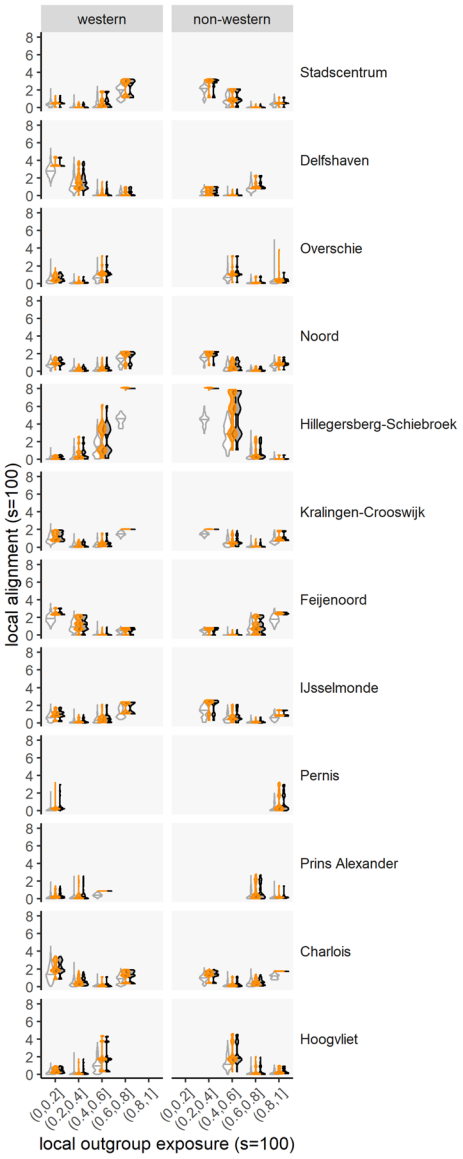

Fig. 19 The effect of outgroup exposure ($s = 100$) on local alignment scores ($s = 100$) for a stratified sample of 500 random agents per simulation run. Alignment scores are measured at the start of the simulation runs (gray violins) and at the end (orange violins). As a reference, black violins represent local alignment scores in case of perfect macro-level alignment. Data are disaggregated by district and ethnic group. Baseline parameter configuration

Acknowledgements The authors are grateful to the members of the research group Norms and Networks at the University of Groningen for their helpful feedback and would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. The third author wishes to acknowledge financial support by the Netherlands Organization for Scientific Research (NWO) under the 2018 ORA grant ToRealSim (464.18.112).

Author contributions *TF*: Conceptualization, Writing–original draft, Investigation, Methodology, Software, Visualization. *JT*: Writing–review & editing, Conceptualization, Investigation, Methodology,

Software, Supervision. *AF*: Conceptualization, Writing–review & editing, Investigation, Methodology, Supervision.

Data availability The datasets generated and analyzed during the current study can be replicated by running the scripts for R 4.2.0 publicly available at https://github.com/thomasfeliciani/scripts-NI-calibration. In order to make these datasets more readily available, the corresponding author further commits to keeping a copy of the generated datasets for at least five years after publication – and to share them upon reasonable request.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Nielsen, M. M., & Hennerdal, P. (2017). Changes in the residential segregation of immigrants in Sweden from 1990 to 2012: using a multi-scalar segregation measure that accounts for the modifiable areal unit problem. *Appl Geog, 87*, 73–84. https://doi.org/10.1016/j.apgeog.2017.08.004
2. Zwiers, M., van Ham, M., & Manley, D. (2018). Trajectories of ethnic neighbourhood change: Spatial patterns of increasing ethnic diversity. *Population Space and Place., 24*(2), e2094. https://doi.org/10.1002/psp.2094
3. Lloyd, C. D., Shuttleworth, I., & Wong, D. W. (2014). Social-spatial segregation: concepts, processes and outcomes. Policy Press.
4. van der Meer, T., & Tolsma, J. (2014). Ethnic diversity and its effects on social cohesion. *Annual Review of Sociology, 40*, 459–478. https://doi.org/10.1146/annurev-soc-071913-043309
5. Bhavnani, R., Donnay, K., Miodownik, D., Mor, M., & Helbing, D. (2014). Group segregation and urban violence. *American Journal of Political Science, 58*(1), 226–245. https://doi.org/10.1111/ajps.12045
6. Weidmann, N. B., & Salehyan, I. (2013). Violence and ethnic segregation: a computational model applied to baghdad. *International Studies Quarterly, 57*(1), 52–64. https://doi.org/10.1111/isqu.12059
7. Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: towards the next frontiers. *Journal of Artificial Societies and Social Simulation, 20*(4), 2. https://doi.org/10.18564/jasss.3521
8. Flache, A. (2019). Social Integration in a Diverse Society: Social Complexity Models of the Link Between Segregation and Opinion Polarization. In F. Abergel (Ed.), *New Perspectives and Challenges in Econophysics and Sociophysics.* Cham: Springer.
9. Feliciani, T., Flache, A., & Tolsma, J. (2017). How, when and where can spatial segregation induce opinion polarization? *Two Competing Models. JASSS, 20*(2), 6. https://doi.org/10.18564/jasss.3419
10. Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology, 21*, 107–112.
11. Festinger, L. (1957). A theory of cognitive dissonance. *Stanford University Press*. https://doi.org/10.1037/10318-001
12. Rosenbaum, M. E. (1986). The repulsion hypothesis: On the nondevelopment of relationships. *Journal of Personality and Social Psychology, 51*(6), 1156–1166. https://doi.org/10.1037/0022-3514.51.6.1156

13. Bovens, M., Jennissen, R., Bokhorst, M., & Engbersen, G. (2021). Afscheid van westers en niet-westers. Naar meervoudige indelingen van herkomstgroepen (WRR-Policy). WRR.
14. Boero, R., & Squazzoni, F. (2005). Does empirical embeddedness matter? methodological issues on agent-based models for analytical social science. *Journal of Artificial Societies and Social Simulation, 8*(4), 6.
15. Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*(3), 7280–7287. https://doi.org/10.1073/pnas.082080899
16. Bruch, E., & Atwell, J. (2015). Agent-based models in empirical social research. *Sociological Methods & Research, 44*(2), 186–221. https://doi.org/10.1177/0049124113506405
17. Heckbert, S., Bishop, I., Marceau, D., & Beneson, I. (2012). Empirical calibration of spatially explicit agent-based models. In D. J. Marceau & I. Benenson (Eds.), *Advanced Geo-simulation Models* (pp. 92–110). Bentham science publishers. USA.
18. Chattoe-Brown, E. (2014). using agent based modelling to integrate data on attitude change. *Sociological Research Online, 19*(1), 16. https://doi.org/10.5153/sro.3315
19. Myers, D. G. (1982). Polarizing Effects of Social Interaction. In H. Brandstätter, J. H. Davis, & G. Stocker-Kreichgauer (Eds.), *Group Decision Making* (pp. 125–161). Academic Press.
20. Vinokur, A., & Burnstein, E. (1978). Depolarization of attitudes in groups. *Journal of Personality and Social Psychology, 36*(8), 872–885. https://doi.org/10.1037/0022-3514.36.8.872
21. Bohner, G., & Dickel, N. (2011). Attitudes and attitude change. *Annual Review of Psychology, 62*(1), 391–417. https://doi.org/10.1146/annurev.psych.121208.131609
22. Asch, S. E. (1955). Opinions and social pressure. *Readings about the Social Animal, 193*, 17–26.
23. Asch, S. E. (1951). A Study of some social factors in perception. *Archives of Psychology, 27*(187), 23–46.
24. Akers, R. L., Krohn, M. D., Lanza-Kaduce, L., & Radosevich, M. (1979). Social learning and deviant behavior: a specific test of a general theory. *American Sociological Review, 44*(4), 636–655.
25. Heider, F. (1967). Attitudes and cognitive organization. In M. Fishbein (Ed.), *Readings in attitude theory and measurement* (pp. 39–41). John Wiley and Sons Inc.
26. Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. Proceedings of the 19th International Conference on World Wide Web 61–70.
27. Balland, P.-A. (2012). Proximity and the evolution of collaboration networks: evidence from research and development projects within the global navigation satellite system (gnss) industry. *Regional Studies, 46*(6), 741–756.
28. Hipp, J. R., & Perrin, A. J. (2009). The simultaneous effect of social distance and physical distance on the formation of neighborhood ties. *City and Community, 8*(1), 5–25.
29. Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M., & Baum, S. (2012). Networks and geography: modelling community network structures as the outcome of both spatial and network processes. *Social Networks, 34*(1), 6–17.
30. Mollenhorst, G., Völker, B., & Schutjens, V. (2009). Neighbour relations in the Netherlands a decade of evidence. *Tijdschrift Voor Economische En Sociale Geografie, 100*(4), 549–558. https://doi.org/10.1111/j.1467-9663.2009.00588.x
31. Massey, D. S., & Denton, N. A. (1988). The Dimensions of Residential Segregation. *Social Forces., 67*(2), 281–315.
32. Sluiter, R., Tolsma, J., & Scheepers, P. (2015). At which geographic scale does ethnic diversity affect intra-neighborhood social capital? *Social Science Research, 54*, 80–95. https://doi.org/10.1016/j.ssresearch.2015.06.015
33. Guiraudon, V., Phalet, K., & ter Wal, J. (2005). Monitoring ethnic minorities in the Netherlands. *International Social Science Journal, 57*(183), 75–87. https://doi.org/10.1111/j.0020-8701.2005.00532.x
34. Reardon, S., & O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology., 650*, 121–162. https://doi.org/10.1111/j.0081-1750.2004.00150.x/abstract
35. Harary, F. (1959). A criterion for unanimity in French's theory of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 168–182). Institute for Social Research.
36. Abelson, R. P. (1964). Mathematical Models of the Distribution of Attitudes Under Controversy. In N. Frederiksen & H. Gulliksen (Eds.), *Contributions to Mathematical Psychology* (pp. 142–160). Rinehart Winston.
37. French, J. R. (1956). A formal theory of social power. *Psychological Review, 63*(3), 181–194.

38. Macy, M. W., Kitts, J. A., Flache, A., & Benard, S. (2003). Polarization in Dynamic Networks: A Hopfield Model of Emergent Structure. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis: workshop summary and papers* (pp. 162–173). The National Academies Press.
39. Jager, W., & Amblard, F. (2005). Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory, 10*(4), 295–303. https://doi.org/10.1007/s10588-005-6282-2
40. Flache, A., & Mäs, M. (2008). Why do faultlines matter? a computational model of how strong demographic faultlines undermine team cohesion. *Simulation Modelling Practice and Theory, 16*(2), 175–191. https://doi.org/10.1016/j.simpat.2007.11.020
41. Musterd, S., & Ostendorf, W. (2009). Residential segregation and integration in the Netherlands. *Journal of Ethnic and Migration Studies, 35*(9), 1515–1532.
42. DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology, 120*(5), 1473–1511. https://doi.org/10.1086/681254
43. Page, S. E. (2015). What sociologists should know about complexity. *Annual Review of Sociology, 41*(1), 21–41. https://doi.org/10.1146/annurev-soc-073014-112230
44. R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/
45. Flache, A., & Macy, M. W. (2011). Small worlds and cultural polarization. *The Journal of Mathematical Sociology, 35*(1–3), 146–176. https://doi.org/10.1080/0022250X.2010.532261
46. Feliciani, T., & Flache, A. (2015). Effects of Groups' Spatial Segregation on Processes of Opinion Polarization. In Paper presented at the Eleventh Conference of the European Social Simulation Association (pp. 1–12).
47. Anselin, L. (1995). Local indicators of spatial association–LISA. *Geographical Analysis*. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x
48. Allport, G. W. (1954). The Nature of Prejudice. *Canadian Psychology*. https://doi.org/10.1037/0708-5591.35.1.11
49. Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751
50. Takács, K., Flache, A., & Mäs, M. (2016). Discrepancy and disliking do not induce negative opinion shifts. *PLoS ONE, 11*(6), e0157948. https://doi.org/10.1371/journal.pone.0157948
51. Keijzer, M. (2022). Opinion dynamics in online social media ICS dissertation series. *Libertas Pascal Utrecht*. https://doi.org/10.33612/diss.196882523
52. Bail, C. A., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: evidence from a large-scale field experiment on social media. SocArXiv Papers. https://doi.org/10.17605/OSF.IO/4YGUX
53. Kozitsin, I. V. (2021). Opinion dynamics of online social network users: a micro-level analysis. *The Journal of mathematical sociology*. https://doi.org/10.1080/0022250x.2021.1956917
54. Liu, C. C., & Srivastava, S. B. (2015). Pulling Closer and Moving Apart: Interaction, Identity, and Influence in the US Senate 1973 to 2009. *American Sociological Review., 80*(1), 192–217. https://doi.org/10.1177/0003122414564182
55. Feliciani, T., Flache, A., & Mäs, M. (2021). Persuasion without polarization? Modelling persuasive argument communication in teams with strong faultlines. *Computational and Mathematical Organization Theory, 27*(1), 61–92. https://doi.org/10.1007/s10588-020-09315-8
56. Mäs, M., & Flache, A. (2013). Differentiation without distancing explaining bi-polarization of opinions without negative influence. *PLoS ONE, 8*(11), e74516. https://doi.org/10.1371/journal.pone.0074516
57. Burger, A., Oz, T., Crooks, A., & Kennedy, W. G. (2017). Generation of Realistic Mega-City Populations and Social Networks for Agent-Based Modeling. Proceedings of the 2017 International Conference of The Computational Social Science Society of the America. Doi: https://doi.org/10.1145/3145574.3145593
58. Crooks, A., Castle, C., & Batty, M. (2008). Key challenges in agent-based modelling for geo-spatial simulation. *Computers, Environment and Urban Systems, 32*(6), 417–430. https://doi.org/10.1016/j.compenvurbsys.2008.09.004
59. Parry, H. R., & Bithell, M. (2012). *Large Scale Agent-Based Modelling: A Review and Guidelines for Model Scaling*. Springer, Netherlands: In Agent-Based Models of Geographical Systems.

60. Macy, M. W., & Willer, R. (2002). From f actors to actors: computational sociology and agent-based modeling. *Annual Review of Sociology, 28*(1), 143–166. https://doi.org/10.1146/annurev.soc.28.110601.141117
61. Flache, A., & De Matos Fernandes, C. A. (2021). *Agent-based computational models*. Edward Elgar Publishing. https://doi.org/10.4337/9781789906851.00033
62. Mäs, M., Flache, A., & Kitts, J. A. (2014). Cultural integration and differentiation in groups and organizations. In V. Dignum & F. Dignum (Eds.), *Perspectives on culture and agent-based simulations* (pp. 71–90). Springer International Publishing. USA.
63. Flache, A., & Mäs, M. (2008). How to get the timing right a computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. *Computational and Mathematical Organization Theory., 14*(1), 23–51. https://doi.org/10.1007/s10588-008-9019-1